U.PORTO
FACULDADE DE CIÊNCIAS
UNIVERSIDADE DO PORTO

MSc

2.º CICLO

FCUP
ANO

U.PORTO

Modelação e Previsão de Velocidade de Ventos

Susana Maria Ferreira Pinto

FC

Modelação e Previsão de Velocidade de Ventos

Susana Maria Ferreira Pinto

Dissertação de Mestrado apresentada à
Faculdade de Ciências da Universidade do Porto
Mestrado em Engenharia Matemática

2015

U.PORTO
FACULDADE DE CIÊNCIAS
UNIVERSIDADE DO PORTO

MSc

Modelação e Previsão de Velocidade de Ventos

Susana Maria Ferreira Pinto
Mestrado em Engenharia Matemática
Departamento de Matemática
2015

Orientadora
Professora Doutora Margarida Brito, FCUP

U.PORTO

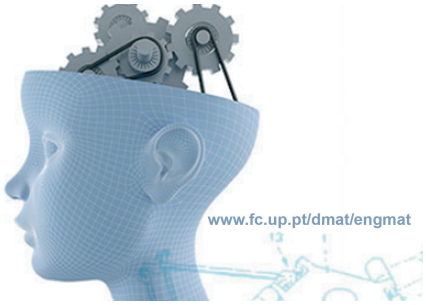
FACULDADE DE CIÊNCIAS
UNIVERSIDADE DO PORTO

Todas as correções determinadas pelo júri, e só essas, foram efetuadas.

O Presidente do Júri,

Porto, _____/_____/_____

Agradecimentos



"Some mathematician, I believe, has said that true pleasure lies not in the discovery of truth, but in the search for it." — *Tolstoi*

Um especial obrigada ...

- Para os meus pais, com quem tenho aprendido a ser quem sou e a quem devo ter chegado até aqui. Por toda a paciência e confiança. Por todos os valores e ensinamentos. Por serem as minhas principais fontes de inspiração.
- Para a minha Orientadora, Professora Doutora Margarida Brito, por ser minha Professora e conselheira e me transformar numa melhor Matemática. Deixo também um agradecimento a todos os meus professores de Licenciatura em Matemática e Mestrado em Engenharia Matemática da Faculdade de Ciências da Universidade do Porto, pelo seu contributo no meu percurso académico.
- Para aqueles que, de uma forma muito especial, marcaram a minha vida pessoal e universitária: Filipe Oliveira, Carla Gonçalves, Luís Baía, Ana Silva, José Pedro Silva, Filipa Carvalho, Júlio Silva, Ana Luísa Lopes, Ricardo Cruz, Marisa Reis e Renato Fernandes. Porque todos nós temos em comum dois aspetos: o gosto pela Matemática e uma panóplia de aventuras para contar.
- Para as amigas de uma vida: Mafalda de Castro, Ana Barros e Joana Freitas. Porque sempre estiveram, estão e estarão lá para mim.
- Para o Aníbal Couto, que me deu as primeiras luzes sobre o mundo dos seguros.
- Para os meus colegas de empresa, por acreditarem em mim e no meu trabalho.

Resumo

Segundo a Organização Mundial de Meteorologia, "é necessário lembrar que mesmo as pequenas obstruções podem causar graves alterações na velocidade do vento e desvios na direção do mesmo".

A determinação precisa da distribuição de probabilidade dos valores de velocidade do vento é muito importante na estimativa de potencial energético da velocidade do vento sobre uma região. A modelação adequada de ventos e estudo da importância de variáveis auxiliares sobre essa modelação (época do ano, existência de precipitação, etc.) pode ser fulcral para entender o que leva, por exemplo, à baixa rentabilidade de uma ventoínha eólica ou até ao possível descarrilamento de um comboio por combinação de ventos fortes e direção adversa do mesmo. A previsão com um grau de certeza considerável permite a sua utilização na meteorologia, aplicação de práticas agrícolas, prevenção de ruína de companhias seguradoras.

Tendo em conta as diferentes aplicações e utilidade deste tipo de estudos, na presente tese são considerados os processos de modelação e previsão de velocidades máximas de vento de uma estação meteorológica italiana. Tem por base o modelo semi-markoviano para a modelação e faz a sua comparação com o modelo markoviano habitualmente usado. Estuda também a previsão da referida variável por árvores de regressão, redes neuronais, regressão com máquinas de suporte vetorial e um modelo linear de séries temporais. Finalmente refere um modelo teórico de modelação de quantidades de ruína de uma seguradora — como referência a uma possível aplicação do processo modelado. A ideia geral é perceber até que ponto é que um processo real — daí o uso de uma base de dados de acesso livre — pode ser explorado com diferentes mecanismos de estudo matemáticos, abordando áreas como *Data Mining*, *Estatística*, *Teoria de Risco*, *Modelação Matemática*, *Processos Estocásticos* e *Simulação Computacional*.

Abstract

According to World Meteorological Organization, "It must always be remembered that even small obstructions cause serious changes in wind speed and deviations in wind direction".

The precise computation of the probability distribution of wind speed values is very important in the estimation of energy potential of wind speed over a region. The proper modelling of wind and study of the importance of auxiliary variables on this modelling (time of year, the presence of precipitation, and so on...) may be crucial to understand what takes, for instance, low profitability of a wind fan or to a possible derailment of a train by a combination of strong winds and its direction. Forecast with a considerable degree of certainty allows its use in meteorology, application of agricultural practices, prevention of ruin of an insurance company.

Taking into account the different applications and utility of such studies, in this thesis are considered processes of modelling and prediction of maximum wind speed of an Italian weather station. It uses the semi-Markovian model for modelling and makes the comparison with the Markovian model commonly used. It also studies the forecast of the variable referred with regression trees, artificial neural networks, regression with support vector machines and a linear time series model. Finally reviews a theoretical model of ruin amounts for an insurance company — reference to a possible application of the modelled process. The general idea is to realize to what extent is that a real process — hence the use of an open access database — can be exploited with different mathematical mechanisms of study, addressing areas such as *Data Mining*, *Statistics*, *Risk Theory*, *Mathematical Modelling*, *Stochastic Processes* and *Computational Simulation*.

Conteúdo

Lista de Tabelas

Lista de Figuras

Capítulo 1

Introdução

A força da Natureza é imparável. Lembrando que os eventos climáticos a ela associados podem ter consequências nefastas (ou, pelo contrário, constituírem um meio para benefício de atividades humanas), torna-se necessário estudá-los, percebendo como é que é possível ***modelá-los*** e ***prevê-los***, uma vez que, quando ocorrem, podem ter grandes impactos materiais e pessoais.

Para tal, foi feito o estudo de uma base de dados que apresenta diversa informação horária registada ao longo de dois anos na estação Meteorológica de Settala (Itália) e que pode ser encontrada no site `http://www.lsi-lastem.com` [1]. O principal foco será o estudo de velocidades máximas de vento. Há diversas aplicações associadas a este tipo de estudo, como por exemplo, produção de energia eólica, definição de práticas agrícolas, previsão meteorológica, entre outras [Aigner e Gjengedal (2011), Chi *et al.* (2007)]. Os processos de modelação e previsão de eventos climáticos têm contribuído para uma melhoria significativa de estratégias de planeamento de atividades frequentemente afetadas pela variabilidade climática, como energia, agricultura e saúde [Green *et al.* (2009)].

A utilização dos resultados provenientes da modelação e previsão de eventos climáticos obriga ao desenvolvimento de técnicas ou métodos que melhorem e aprimorem este tipo de trabalho. Além disso, é de conhecimento geral que a realidade muitas vezes não se assemelha às simplificações da teoria e que, por isso, as simplificações teóricas podem ser demasiadas para conseguir modelar adequadamente um determinado evento. Do mesmo modo, é comum ver exemplos de dados muito bem comportados para os quais os modelos funcionam muito bem. E com dados reais, é também esse o comportamento que se deve esperar? Portanto existe, nesta tese, não só uma introdução de técnicas usadas para os diferentes objetivos traçados anteriormente, mas também uma aplicação dos mesmos a dados reais.

Foi realizada uma separação física da informação, neste documento, em três capítulos fundamentais: os dois primeiros fazem a contextualização teórica inicialmente pensada para o caso prático, definido no capítulo seguinte.

[1] última consulta realizada em 28 de Junho de 2015.

Capítulo 2

Modelação

"... all models are approximations. Essentially, all models are wrong, but some are useful. However, the approximate nature of the model must always be born in mind..." — George E.P. Box

As cadeias de Markov [Ross (2014)] são o processo mais utilizado para modelação dos ventos. Uma das grandes dificuldades que a modelação de ventos evidencia, quando feita desta forma, é a falta de flexibilidade que as cadeias de Markov apresentam em relação aos tempos de espera, o que faz com que as simulações dos ventos sejam pouco fiéis aos dados originais. Os processos semi-markovianos têm sido amplamente utilizados para modelar fenómenos naturais [Asaduzzaman e MahbubLatif (2013), Barbu *et al.* (2004), Sansom *et al.* (2001)] e formam uma classe de processos estocásticos que generalizam, em simultâneo, cadeias de Markov e processos de renovamento. A principal vantagem da sua utilização, quando comparados com os processos Markovianos, é o uso de qualquer distribuição para modelar os tempos de espera.

2.1 Breve revisão de processos de Markov

Para contextualizar o problema, vai ser feita uma breve revisão de processos estocásticos. Um processo estocástico é uma família ou conjunto de variáveis aleatórias definidas num espaço de probabilidade - e indexadas no tempo ou no espaço.

Assuma-se que $\{X(t)\}$ é um processo aleatório indexado no tempo $t \in T$ e $X(t)$ é o estado do processo. Se o conjunto T é finito ou infinito numerável, então $\{X(t)\}$ é um processo em tempo discreto. Caso contrário, é um processo em tempo contínuo. Do mesmo modo, se $\{X(t)\}$ é definido sob um conjunto contável de estados, então o processo é definido por uma cadeia, uma vez que é discreto nos estados. Caso contrário, é contínuo nos estados e designado como uma sequência.

Considerando $x(t)$ como as realizações de $X(t)$ e assumindo que $x(t) \in \mathbb{R}$, podem definir-se as seguintes funções:

- média como função do tempo: $\mu_X(t) = m_X(t) = E[X(t)] = \int_{-\infty}^{\infty} x f_{X(t)}(x(t))dx$ onde $f_{X(t)}$ representa a função densidade de probabilidade de $X(t)$
- autocorrelação de segunda ordem:

$$R_X(t_1, t_2) = E[X(t_1)X(t_2)] = \int_{-\infty}^{\infty} x_1 x_2 f_{X(t_1),X(t_2)}(x_1(t_1), x_2(t_2))dx_1 dx_2$$

- autocovariância: $C_X(t_1, t_2) = Cov(X(t_1), X(t_2)) = R_X(t_1, t_2) - \mu_X(t_1)\mu_X(t_2)$ (observe-se que $Var(X(t)) = C_X(t, t)$)

Sob um ponto de vista de classificação, e definindo a função de distribuição conjunta como

$$F_{X_1(t_1),...,X_n(t_n)}(x_1, ..., x_n) = P(X(t_1) \leq x_1, ..., X(t_n) \leq x_n),$$

um processo estocástico pode ser classificado como:

- **estacionário de ordem** n se os momentos de ordem $\leq n$ são independentes do tempo;
- **estacionário em sentido estrito** se, $\forall n, \forall\{t_1, t_2, ..., t_n\}, \forall\tau \in \mathbb{R}$

$$F_{X_1(t_1),...,X_n(t_n)}(x_1, ..., x_n) = F_{X_1(t_1+\tau),...,X_n(t_n+\tau)}(x_1, ..., x_n);$$

- **estacionário em sentido lato** se

$$\mu_X(t) = E(X(t)) = \mu_X \quad \text{e} \quad R_X(t + \tau, \tau) = R_X(\tau) \quad \tau = t_1 - t_2$$

  ou seja, se a média for constante ao longo do tempo e a função de autocorrelação só depender da diferença entre t_2 e t_1. Note-se que um processo estacionário em sentido lato não implica que o mesmo o seja em sentido estrito, uma vez que a primeira propriedade é menos exigente que a segunda.

Seja $\{N(t)\}$ o **processo de contagem** do número de eventos no intervalo $(0, t]$ e

$$N(0) = 0 \leq N(t_1) \leq N(t_2) \leq ... \leq N(t_k) \leq ... \qquad \forall\ 0 \leq t_1 \leq ... \leq t_k \leq ...$$

Lembrando que $N(t) = \sum_{n \geq t} I_{\{T_n \leq t\}}$, onde $I_{\{T_n \leq t\}} = \begin{cases} 1 & \text{se } T_n \leq t \\ 0 & \text{se } T_n > t \end{cases}$, note-se que

$$\{T_n, n \in \mathbb{N}\} \equiv \{N(t), t \geq 0\}$$

porque $N(t) = n \Leftrightarrow T_n \leq t < T_{n+1}$ e $N(t) \geq n \Leftrightarrow T_n \leq t$. Esta dualidade é muito útil, porque permite escrever um problema à custa do outro, característica frequentemente utilizada no estudo deste tipo de processos.

Diz-se que o processo de contagem, $\{N(t), t \geq 0\}$ tem:

- incrementos independentes se

$$N(0), N(t_1) - N(0), N(t_2) - N(t_1), ..., N(t_n) - N(t_{n-1})$$

são variáveis aleatórias independentes, o que significa que o número de eventos ocorridos até ao instante t é independente do número de eventos que ocorrem entre t e $t+s$, para algum s.

- incrementos estacionários, se $P(N(t+s) - N(t) = k) = P(N(s) = k)$ para qualquer $t \geq 0$.

Definição: Diz-se que um processo $\{N(t), t \geq 0\}$ é de Poisson se tem incrementos estacionários e independentes, $N(0) = 0$, $P(N(h) = 1) = \lambda h + o(h)$ e $P(N(h) \geq 2) = o(h)$.

Proposição: Se $\{N(t), t \geq 0\}$ é um processo de Poisson, então $N(t) \sim Poisson(\lambda t)$, onde λ é o parâmetro correspondente à intensidade do processo.

Assumindo que $\{N(t)\}$ segue um processo de Poisson, a distribuição dos tempos entre eventos de uma cadeia de Markov tem distribuição exponencial (a explicação desta afirmação pode ser encontrada no Anexo A).

Um **processo de Markov** é um processo estocástico caracterizado pela *propriedade markoviana*, segundo a qual, a probabilidade de qualquer comportamento futuro do processo, quando o seu estado atual é conhecido, não é alterada pela existência de conhecimento adicional sobre o seu comportamento passado. Esta informação é traduzida pela seguinte expressão:

$$P(X(t_{k+1}) \leq x_{k+1} \mid X(t_k) = x_k, ..., X(t_0) = x_0) = P(X(t_{t+1}) \leq x_{k+1} \mid X(t_k) = x_k).$$

2.2 Introdução aos processos Semi-markovianos

Um processo estocástico de Semi-Markov é uma generalização de um processo de Markov, uma vez que a informação sobre o tempo de permanência no estado atual passa a ser relevante. Contudo, mantém-se irrelevante para o comportamento futuro qualquer informação sobre os estados visitados no passado. Consequentemente, os tempos entre acontecimentos sucessivos deixam de estar restritos à distribuição exponencial, podendo seguir qualquer distribuição de probabilidade. Seguidamente apresentar-se-ão os conceitos fundamentais para a definição e compreensão de um processo deste tipo, mas uma introdução mais detalhada pode ser encontrada, por exemplo, em Iosifescu *et al.* (2013) e Janssen e Manca (2006).

Considere-se $I = \{1, ..., M\}$ como um espaço finito de estados e (Ω,F,P) como um espaço de probabilidade. Sejam

$$J_n : \Omega \to I \quad \text{e} \quad T_n : \Omega \to \mathbb{N}$$

duas variáveis aleatórias, sendo que J_n representa o estado ocupado pelo sistema imediatamente após ter efetuado a n-ésima transição e T_n representa o instante de tempo em que ocorreu a n-ésima transição ($n \in \mathbb{N}$). $X_n = T_n - T_{n-1}$ é o tempo que o sistema levou a transitar de J_{n-1} para J_n. Deste modo, $(X_n)_{n\in\mathbb{N}}$ designa uma sequência de tempos de espera e $(T_n)_{n\in\mathbb{N}}$ uma sequência de tempos de chegada.

Note-se que $N(t) = sup\{n : T_n \leq t\} \quad \forall t \in \mathbb{N}$ representa o número de transições que ocorreram em $(0, t]$.

Por exemplo, considerando $M = 2$, uma possível realização do processo estocástico bidimensional (J_n, X_n) seria aquela em que o sistema começa no estado $J_0 = 2$, transita para 1 após 2 unidades de tempo ($J_1 = 1$, $X_1 = 2$) e transita novamente para 2 após 3 unidades de tempo ($J_2 = 2$, $X_2 = 3$).

O par (J_n, T_n) designa o **processo Markoviano não-homogéneo de renovamento**. Note-se que os processos de renovamento fornecem modelos teóricos de investigação para a ocorrência de padrões em experiências independentes e repetidas [Pyke (1961)]. De um modo informal, pode escrever-se que o termo *renovamento* vem do pressuposto básico de que, quando o padrão de interesse ocorre pela primeira vez, o processo se repete, no sentido em que a situação inicial é restabelecida.

Introduzidos os conceitos básicos necessários para a compreensão de um modelo semi-markoviano, introduz-se o núcleo não-homogéneo semi-markoviano associado a este processo, $Q = [Q_{ij}(s, t)]$, que representa a probabilidade de chegada a um determinado estado até um determinado instante, sabendo qual o estado e instante da transição anterior, e é definido como

$$\begin{aligned} Q_{ij}(s,t) &= P(J_{n+1} = j, X_{n+1} \leq t - s \mid J_n = i, T_n = s) \\ &= P(J_{n+1} = j, T_{n+1} \leq t \mid J_n = i, T_n = s) \end{aligned}$$

portanto $P = [p_{ij}(s)]$ define a matriz de transições, onde

$$p_{ij}(s) = P(J_{n+1} = j \mid J_n = i, T_n = s)$$

devolve a probabilidade do sistema estar no estado j sabendo que no instante s estava no estado i. Note-se ainda que

$$p_{ij}(s) = \lim_{t\to\infty} Q_{ij}(s, t) \quad i, j \in I; s, t \in \mathbb{N}, s \leq t.$$

Para além do núcleo Q, existem outras probabilidades condicionadas relevantes para a definição de um processo semi-markoviano. Uma função particularmente relacionada com a construção do núcleo é a função de distribuição condicional do tempo de espera em cada estado i, dado o estado subsequente ocupado:

$$F_{ij}(s, t) = P(X_{n+1} \leq t - s \mid J_n = i, J_{n+1} = j, T_n = s)$$

$$= P(T_{n+1} \leq t \mid J_n = i, J_{n+1} = j, T_n = s) = \begin{cases} \frac{Q_{ij}(s,t)}{p_{ij}(s)} & \text{se } p_{ij}(s) \neq 0 \\ 1 & \text{se } p_{ij}(s) = 0 \end{cases}$$

Note-se que esta função pressupõe que uma certa transição vai, de facto, ocorrer, devolvendo apenas a probabilidade de ocorrer numa certa duração (ou, de modo equivalente, até um determinado instante de tempo).

Uma vez que o núcleo markoviano é a função responsável pela produção de qualquer resultado deste modelo (daí a sua designação), é comum utilizar P e F para a construção do núcleo Q, fazendo uso da relação vista anteriormente.

Com a finalidade de simplificar a construção do modelo, considerem-se as funções que definem:

- A probabilidade de que o processo deixe um determinado estado i até ao instante t, sabendo que no instante s tinha chegado ao estado i:

$$H_i(s,t) = P(T_{n+1} \leq t \mid J_n = i, T_n = s) = \sum_{j=1}^{M} Q_{ij}(s,t).$$

- a probabilidade de que o sistema chegue ao estado j no instante t, sabendo que no instante s tinha chegado ao estado i:

$$b_{ij}(s,t) = P(J_{n+1} = j, T_{n+1} = t \mid J_n = i, T_n = s).$$

  Note-se que esta probabilidade apenas pode ser definida quando se assume que o processo segue um percurso temporal discreto e que, por sua vez, pode ser escrita em função de $Q_{ij}(s,t)$:

$$b_{ij}(s,t) = \begin{cases} Q_{ij}(s,t) - Q_{ij}(s,t-1) & \text{se } s < t \\ 0 & \text{se } s = t \end{cases}$$

Estão agora definidas as ferramentas necessárias para caracterizar um dos principais outputs deste modelo, e uma das suas funções de interesse: o **processo semi-markoviano de primeira ordem, não-homogéneo em tempo discreto $\boldsymbol{Z = (Z(t))}$**, que representa, para cada instante de tempo t, o estado ocupado pelo processo. As probabilidades de transição para este processo são dadas por

$$\phi_{ij}(s,t) = P(Z(t) = j \mid Z(s) = i).$$

Note-se que $Z(t)$ e $J_{N(t)}$ são o mesmo processo. No entanto, $\phi_{ij}(s,t)$ difere de $Q_{ij}(s,t)$, na medida em que o primeiro considera a possibilidade de transições intermédias até à chegada ao estado j, ao passo que o segundo considera a probabilidade de chegada a esse estado na transição seguinte.

As probabilidades de transição do processo $Z(t)$ são definidas como

$$\phi_{ij}(s,t) = \delta_{ij}(1 - H_i(s,t)) + \sum_{\beta=1}^{M} \sum_{\vartheta=s}^{t} b_{i\beta}(s,\vartheta)\phi_{\beta j}(\vartheta,t) \quad \text{[Janssen e Manca (2002)]}. \quad (2.1)$$

onde δ_{ij} representa o delta de Kronecker, dado por $\delta_{ij} = \begin{cases} 1 & \text{se } i = j \\ 0 & \text{se } i \neq j \end{cases}$. O resultado anterior foi provado pelos mesmos autores, que também demonstraram que esta equação admite uma solução única [Janssen e Manca (2006)]. Para uma maior intuição do conceito, deve interpretar-se a expressão anterior: o primeiro termo representa a probabilidade de permanência do processo no estado i durante os instantes s e t (porque o percurso do processo não é pré-definido e não existe garantia de que transite de estado) e o segundo termo representa a probabilidade de transição do processo do estado i para o estado j entre os instantes s e t (notando que entretanto pode passar por outros estados β em instantes intermédios ϑ). Como esta é uma função diferencial que depende de $H_i(s,t)$ e de $b_{i\beta}(s,\vartheta)$, basta a definição de uma condição de fronteira para que se possa resolver. Quando $s = t$, e como no mesmo instante de tempo não se pode estar em dois estados diferentes, a condição referida será dada por

$$\phi_{ij}(s,s) = \begin{cases} 1 & \text{se } i = j \\ 0 & \text{se } i \neq j \end{cases}$$

e a equação pode ser resolvida do seguinte modo:

- Define-se um valor máximo para o tempo que se vai estudar, $T_{\max}$.
  Define-se $\phi_{ij}(s,s) \quad \forall s \in 0, ..., T_{\max}$;
- Calcula-se $\phi_{ij}(T_{\max} - 1, T_{\max})$. Analisando a equação que define o processo, verifica-se que este é um cálculo trivial, uma vez que só depende de $\phi_{ij}(T_{\max}, T_{\max})$, conhecido pelo passo anterior;
- Calcula-se $\phi_{ij}(T_{\max} - 2, T_{\max})$, que necessita de $\phi_{ij}(T_{\max} - 1, T_{\max})$ e $\phi_{ij}(T_{\max}, T_{\max})$, ambos valores já obtidos;
- Calcula-se $\phi_{ij}(T_{\max} - 3, T_{\max})$, que depende de $\phi_{ij}(T_{\max} - 1, T_{\max})$, $\phi_{ij}(T_{\max}, T_{\max})$ e $\phi_{ij}(T_{\max} - 2, T_{\max})$, entretanto conhecidos;
- Continua-se o processo até que $\phi_{ij}(0, T_{\max})$ esteja calculado;
- Repete-se o processo, iniciando em $\phi_{ij}(T_{\max} - 1, T_{\max-1})$.

Uma vez resolvida esta equação, será possível analisar todas as probabilidades de transição e permanência do sistema Z.

Lembrando a definição de $N(t)$, seja ainda

$$B(t) = t - T_{N(t)}, \quad t \in \mathbb{N}$$

o tempo desde a última transição, podendo ser interpretado como o processo temporal recursivo de $Z(t)$ (*backward process*). Notando que (Z, B) é um processo de Markov, é agora

possível escrever as diferentes funções vistas anteriormente, mas tendo agora em consideração este novo processo B:

$$ {}^{b}H_i(u,s,t) = P(T_{N(s)+1} \leq t \mid J_{N(s)} = i, T_{N(s)} = u, T_{N(s)+1} > s) \quad (u \leq s < t) $$

$$ {}^{b}Q_{i,j}(u,s,t) = P(T_{N(s)+1} \leq t, J_{N(s)+1} = j \mid J_{N(s)} = i, T_{N(s)} = u, T_{N(s)+1} > s) \quad (u \leq s < t) $$

Observando que ${}^{b}H_i(s,s,t) = H_i(s,t)$ e ${}^{b}Q_{i,j}(s,s,t) = Q_{i,j}(s,t)$, conclui-se que

$$ {}^{b}Q_{i,j}(u,s,t) = \frac{{}^{b}Q_{i,j}(u,u,t)}{1 - {}^{b}H_i(u,u,t)} = \frac{Q_{i,j}(u,t)}{1 - H_i(u,s)} $$

Portanto ${}^{b}\phi_{i,j}(u,s,t) = P(Z(t) = j \mid T_{N(s)} = u, Z[u,s] = i) =$

$$ = P(Z(t) = j \mid Z(s) = i, B(s) = s - u) $$

devolve a probabilidade, sabendo que o sistema no instante de tempo s estava no estado i e que entrou nesse estado no instante u (por outras palavras, que estava no estado i há $s - u$ unidades de tempo), de estar no instante t no estado j. Este é um resultado com muito potencial. Por exemplo, a probabilidade do sistema se encontrar num estado de vento considerado grave poderá depender de há quanto tempo é que se encontra num estado com essa categorização.

Para construir qualquer output do modelo é necessário estimar Q. Há duas formas de o fazer:

1 Estimar as funções $p_{ij}(s)$ e $F_{ij}(s,t)$. Foque-se a **forma empírica:** definir $p_{ij}(s)$ como a razão entre o número de observações que se encontravam no estado i no instante s cujo estado seguinte foi j e o número de observações que se encontravam no estado i no instante s. Considerando a_{ij} como o número de transições que ocorreram do estado i para o estado j, Barbu e Limnios (2009) mostram que o estimador empírico $\hat{p}_{ij} = \frac{a_{ij}}{\sum_{j=1}^{M} a_{ij}}$ se aproxima do estimador por máxima verosimilhança. Será este o utilizado nas simulações.

Analogamente, definir empiricamente $F_{ij}(s,t)$ como a razão entre o número de observações que se encontravam no estado i no instante s, cujo estado seguinte foi j e cuja transição aconteceu no instante t e o número de observações que se encontravam no estado i no instante s, cujo estado seguinte foi j. Observe-se ainda que $p_{ij}(s)$ não tem necessariamente de depender de s. Pode assumir-se que $p_{ij}(s) = p_{ij}(0) \quad \forall s$, pelo que a probabilidade de transição entre estados seria a mesma ao longo do tempo.

2 Estimar diretamente $Q_{ij}(s,t)$. Para a estimação direta de Q, note-se que a estimação empírica consistiria na análise da probabilidade de haver uma transição no preciso instante de tempo em estudo, e fazer essa análise para todos os instantes de tempo considerados. Por ser menos intuitiva, apesar de exequível, analisar-se-á apenas o caso anterior.

No entanto, o modelo anteriormente definido pode ser ainda mais abrangente. Um exemplo é a abordagem feita em D'Amico *et al.* (2013), que sugere o uso do modelo semi-markoviano de segunda ordem no espaço de estados e de tempos. Para referir os conceitos principais e simplificados deste modelo, vão utilizar-se notações similares às anteriores. No modelo semi-markoviano de segunda ordem sugerido nesse artigo:

- É assumida a seguinte relação:

$$P(J_{n+1}, T_{n+1} - T_n = t \mid \sigma(J_s, T_s), J_n = k, J_{n-1} = i, T_n - T_{n-1} = x, 0 \leq s \leq n)$$

$$= P(J_{n+1}, T_{n+1} - T_n = t \mid J_n = k, J_{n-1} = i, T_n - T_{n-1} = x)$$

  (onde $\sigma(J_s, T_s)$ representa o conjunto possível de estados e de tempos da transição s estudada - esta é uma mera formalidade de escrita usada para garantir o conhecimento dos valores possíveis das grandezas estudadas) que garante que o conhecimento dos estados J_n e J_{n-1}, assim como dos tempos T_n e T_{n+1} é suficiente para construir a distribuição condicional de J_{n+1} e T_{n+1};

- ${}_xQ_{i,k,j}(t) = P(J_{n+1} = j, T_{n+1} - T_n \leq t \mid J_n = k, J_{n-1} = i, T_n - T_{n-1} = x)$ representa o núcleo semi-markoviano de segunda ordem;

- ${}_xF_{i,k,j}(t) = P(T_{n+1} - T_n \leq t \mid J_n = k, J_{n-1} = i, J_{n+1} = j, T_n - T_{n-1} = x)$ representa a distribuição condicional dos tempos de permanência, dado o estado subsequente ocupado;

- A função que modela a distribuição dos tempos de permanência nos estados i e k com duração x é dada por ${}_xH_{i,k}(t) = P(T_{n+1} - T_n \leq t \mid J_n = k, J_{n-1} = i, T_n - T_{n-1} = x)$.

Dadas estas definições, a cadeia semi-markoviana neste modelo seria

$$Z(t) = (Z^1(t), Z^2(t)) = (J_{N(t)-1}, J_{N(t)}).$$

As simplificações e detalhes da implementação num caso prático serão explicados no capítulo de Aplicação.

Capítulo 3

Previsão

"Prediction is very difficult, especially if it's about the future." — Niels Bohr

O vento é, ainda hoje, considerado uma das variáveis de clima mais difíceis de ser prevista. No caso da base de dados de que se dispõe, o objetivo será fazer a previsão de velocidade do vento tendo por base a série temporal do mesmo. Recorrendo à informação temporal que existe, o principal objetivo deste capítulo foi perceber até que ponto é que é possível prever quais as velocidades de vento futuras naquela região de Itália e, se for possível, fazê-lo.

A estratégia para previsão de dados, quaisquer que sejam, deve compreender as etapas de *identificação do modelo*, *estimação de parâmetros* e *verificação de diagnóstico*. Estas etapas são repetidas iterativamente até ser encontrado um modelo satisfatório para os dados.

3.1 Previsão - Séries Temporais & Modelos ARIMA

Uma série temporal pode ser definida como um conjunto de observações de uma variável realizadas em períodos sucessivos de tempo. São usualmente analisadas a partir de características como tendência, ciclo, sazonalidade e variações aleatórias e existem diversas famílias de modelos propostos na literatura para a análise de séries temporais [Gonçalves e Lopes Mendes (2008)].

A principal ideia das séries temporais é a suposição de que o valor seguinte de uma série se pode explicar a partir dos valores anteriores e da influência de perturbações aleatórias, formando uma série de ruído branco. As diferentes combinações destas influências resultam em diferentes formatos de modelos, cuja construção depende de fatores, como o comportamento esperado do fenómeno ou o conhecimento *a priori* que se tem sobre ele.

Existem dois indicadores estatísticos dos dados, frequentemente usados, que permitem inferir alguma informação sobre a autocorrelação dos mesmos, constituíndo uma ferramenta de decisão sobre que modelo de previsão se deve usar para a série temporal em estudo:

- Função de autocorrelação **ACF**
- Função parcial de autocorrelação **PACF**

A ACF é uma função que avalia a correlação entre a série estudada e ela própria desfasada de k períodos de tempo. É obtida pela expressão

$$\rho_k = \frac{\gamma_k}{\sigma^2} = \frac{E[(Y_t - \bar{Y})(Y_{t+k} - \bar{Y})]}{\sigma^2}, \quad k = 1, ..., n-1$$

onde Y_t corresponde ao valor observado no tempo t, $\bar{Y}$ ao valor médio da série temporal, σ^2 à variância e n ao número de observações disponíveis. A PACF devolve os coeficientes de correlação parcial entre a série e ela própria desfasada de k períodos de tempo e é obtida pela expressão

$${}^{\mathrm{P}}\rho_\tau = \frac{\rho_\tau - \sum_{j=1}^{\tau-1} {}^{\mathrm{P}}\rho_{\tau-1,j}\rho_{\tau-j}}{1 - \sum_{j=1}^{\tau-1} {}^{\mathrm{P}}\rho_{\tau-1,j}\rho_{\tau-j}}, \quad \text{para } \tau > 1.$$

Portanto, enquanto que a autocorrelação de ordem k mede a dependência linear entre Y_{t-k} e Y_t, a autocorrelação parcial de ordem k é a correlação existente entre os resíduos de Y_t e Y_{t-k}, após efectuar a regressão linear de cada uma destas variáveis sobre Y_{t-1},...,Y_{t-k+1}.

Um modelo frequentemente utilizado para fazer previsão em séries temporais é o modelo ARIMA (*Autoregressive Integrated Moving Average*), que depende de três parâmetros: p, d, q. O parâmetro p é referente à parte auto-regressiva, o parâmetro d diz respeito ao número de diferenciações que são necessárias para transformar a série não estacionária numa série estacionária e o parâmetro q determina o número de médias móveis a ser usado. Para se aplicar um modelo deste tipo, devem seguir-se os seguintes passos:

1. Caso a série temporal não seja estacionária, transformá-la para que o seja (em sentido lato); para o fazer, pode diferenciar-se a série d vezes. Observe-se que, pelo termo "diferenciar", se assume a transformação dada pelas diferenças sucessivas da série original (isto é, se Y_t for a série temporal, então $\triangle Y_t = Y_t - Y_{t-1}$ será a primeira diferenciação, $\triangle^2 Y_t = \triangle[\triangle Y_t] = Y_t - 2Y_{t-1} + Y_{t-2}$ a segunda, ...);

2. Estimar os parâmetros p e q.

Estes modelos foram introduzidos por Box e Jenkins (1970) e têm a vantagem de serem bastante flexíveis, na medida em que podem representar diferentes tipos de séries temporais, separando-se em auto-regressivo puro (AR), médias móveis puro (MA) e série combinada ARMA. No entanto, assumem uma estrutura de correlação linear entre os valores da série temporal, pelo que os padrões não-lineares podem ser capturados pelos resíduos do modelo.

Note-se que um modelo é designado por autoregressivo e de médias móveis quando é possível escrever a série temporal a partir dos seus valores passados e de eventuais perturbações aleatórias, respetivamente.

Formalmente:

- Um processo estocástico centrado $Y = (Y_t, t \in \mathbb{Z})$ de segunda ordem, estacionário ou não, admite uma representação auto-regressiva de ordem p se existem números reais $\varphi_1, \varphi_2, ..., \varphi_p$ e um ruído branco $\varepsilon_t, t \in \mathbb{Z}$ de variância σ^2 ($\sigma^2 > 0$) tais que

$$Y_t - \varphi_1 Y_{t-1} - ... - \varphi_p Y_{t-p} = \varepsilon_t \quad \text{com } \varepsilon \neq 0.$$

Note-se que uma sequência $\{\varepsilon_t\}$ é dita **ruído branco** se cada valor da série tiver média zero, variância constante e não apresentar correlação serial.

- Um processo de segunda ordem $Y = (Y_t, t \in \mathbb{Z})$ admite uma representação média de ordem q se
$$Y_t = \varepsilon_t - \theta_1\varepsilon_{t-1} - \theta_2\varepsilon_{t-2} - ... - \theta_q\varepsilon_{t-q} \quad \forall t \in \mathbb{Z}$$

- Um processo $Y = (Y_t, t \in \mathbb{Z})$ que admita uma representação auto-regressiva média móvel verifica uma equação da forma
$$Y_t - \varphi_{t-1}Y_{t-1} - ... - \varphi pY_{t-p} = \varepsilon_t - \theta_1\varepsilon_{t-1} - \theta_2\varepsilon_{t-2} - ... - \theta_q\varepsilon_{t-q}.$$

Note-se que os recursos de estudo de séries temporais foram elaborados sobretudo para séries estacionárias. Portanto é necessário perceber como se pode testar a estacionariedade de uma série temporal. Uma das formas de o fazer é testando a existência de alguma raíz unitária com base nas seguintes hipóteses:

$$H_0 = \text{“Existe pelo menos uma raiz dentro do círculo unitário”}$$
$$H_1 = \text{“Não existem raízes dentro do círculo unitário”}.$$

Um processo estocástico linear tem uma raíz unitária se 1 é raíz da equação característica[1]. Um processo com essa característica é **não estacionário**. Se as outras raízes da equação característica estão dentro do círculo unitário, isto é, têm valor absoluto inferior a 1, a primeira diferenciação do processo será estacionária. Por esse motivo, é comum encontrar testes de hipóteses associados à raíz unitária de um determinado processo estocástico, para testar a estacionariedade de uma série temporal. Os mais utilizados e que vão ser aplicados à base de dados estudada, são explicados agora:

1. **Dickey-Fuller Aumentado** (*Augmented Dickey-Fuller*)**:** Este teste requer o estudo da regressão
$$\Delta y_t = \beta_1 + \beta_2 t + \delta y_{t-1} + \sum_{i=1}^{m} \alpha_i \Delta y_{t-i} + \epsilon_t$$
onde β_1 é a interceção, β_2 é o coeficiente de tendência, δ é o coeficiente de presença de raiz unitária e m é o número de desfasamentos considerados. A hipótese nula é dada por $\delta = 0$. Faz-se uma regressão de Δy_t em $y_{t-1}, \Delta y_{t-1}, \ldots, \Delta y_{t+p-1}$ e calcula-se a estatística de teste dada por
$$\frac{\hat{\delta}}{\hat{\sigma}(\delta)}$$
onde $\hat{\delta}$ é um estimador para δ e $\hat{\sigma}(\delta)$ é um estimador para o desvio-padrão do erro de δ.

[1] A equação característica é diferente consoante o processo analisado. Pode considerar-se, no caso de um modelo AR(p), o polinómio autoregressivo f de ordem p e a equação característica será dada por $f(L) = 0$, onde L é o operador de diferenciação (ou *lag*) do processo; no caso de um modelo MA(q), o polinómio g de médias móveis de ordem q, cuja equação característica será dada por $g(L) = 0$. Mais informação pode ser encontrada em Gonçalves e Lopes Mendes (2008).

O detalhe da proposta original deste teste pode ser consultado em Dickey e Fuller (1979).

2. **Phillips-Perron:** Este teste foi formulado com o objetivo de resolver o viés assintótico do teste original de Dickey-Fuller, nos casos em que existe correlação entre os resíduos [Davidson e MacKinnon (1993)]. O teste de Phillips-Perron faz uma correção não paramétrica para a estatística de teste, estudando a mesma regressão e hipótese nula vistas anteriormente, mas permitindo a sua utilização na presença de variáveis desfasadas dependentes e correlação entre os resíduos da série.

   Neste teste, a estatística a usar será dada por

$$n\hat{\delta_n} - \frac{n^2\hat{\sigma}^2}{2{s^2}_n}(\hat{\lambda}_n^2) - \hat{\gamma}_{0,n}$$

   considerando

$$\hat{\gamma}_{j,n} = \frac{1}{n}\sum_{i=1+j}^{n} r_i r_{i-j}, \quad \hat{\lambda}_n^2 = \hat{\gamma}_{0,n} + 2\sum_{j=1}^{q}(1 - \frac{j}{q+1})\hat{\gamma}_{j,n} \quad \text{e} \quad {s^2}_n = \frac{1}{n-k}\sum_{i=1}^{n} {r_i}^2,$$

   onde r_i representa o resíduo em y_i, k o número de covariáveis na regressão e q o número de desfasamentos temporais usados em $\hat{\lambda}_n^2$.

   No entanto, Davidson e MacKinnon (2004) relatam que o teste de Phillips-Perron tem um pior desempenho em amostras finitas do que o teste de Dickey-Fuller aumentado.

3. **KPSS:** Teste proposto por Denis Kwiatkowski , Peter C. B. Phillips, Peter Schmidt e Yongcheol Shin [Kwiatkowski *et al.* (1992)]. As hipóteses a testar são:

$$H_0 = \text{“A série é estacionária”}$$
$$H_1 = \text{“A série apresenta raiz unitária”}.$$

   Considere-se

$$Y_t = \beta t + r_t + \epsilon_t,$$

   e assuma-se que a série admite decomposição em componentes de tendência, passeio aleatório e erro, com $r_t = r_{t-1} + \mu_t$ (onde μ_t i.i.d, com média zero e variância σ_μ^2) e resíduos $(\epsilon_t)_{t=1,2,\ldots,T}$ (onde T é o instante temporal máximo avaliado). A estatística de teste é dada por

$$\frac{\sum_t S_t^2}{T^2\hat{\sigma}_{\epsilon_t}^2}$$

   denotando $\hat{\sigma}_{\epsilon_t}^2$ como estimador para a variância dos erros nesta regressão e $S_t = \sum_{i=1}^{t}\epsilon_i$ como a soma parcial dos resíduos. A estatística considerada tem valores críticos tabelados.

Sob ponto de vista analítico, a ideia será aplicar os testes em conjunto para avaliar a estacionariedade da série.

3.2 Previsão - Métodos de *Data Mining*

Inserida num contexto de processamento computacional e de procura de padrões em grandes conjuntos de dados, esta secção visa apresentar previsão através de diferentes métodos vulgarmente usados em *Data Mining* [Hand *et al.* (2001)]. É importante lembrar que a previsão apenas é possível se se assumir que existe alguma regularidade nas observações do evento a prever. Os problemas de previsão separam-se em **classificação** - caso a variável a prever seja nominal - e **regressão** - caso a variável a prever seja numérica. Existem várias técnicas que podem ser usadas para previsão, mas todas têm em comum algumas características:

- Suposição de uma forma funcional para a função que explica a variável a prever, com base nas variáveis que a descrevem (ou explicam);
- Definição de um critério de decisão para a escolha do melhor modelo de previsão. O critério preferencial é o da **minimização do erro de previsão**.

As técnicas utilizadas serão as seguintes:

- Redes Neuronais Artificiais;
- Árvores de Regressão;
- Regressão com Máquinas de Suporte Vetorial.

3.2.1 Redes Neuronais Artificiais

Este método faz uma analogia com o funcionamento das redes neuronais do sistema nervoso (como ilustrado na Figura 3.1), sendo por isso um modelo inspirado nesta estrutura de unidades celulares que adquirem conhecimento através da experiência. As ANN (*Artificial Neural Network*) são frequentemente utilizadas em reconhecimento de padrões, classificação de séries temporais e aproximação de funções, entre outras aplicações [LeCun *et al.* (1998)].

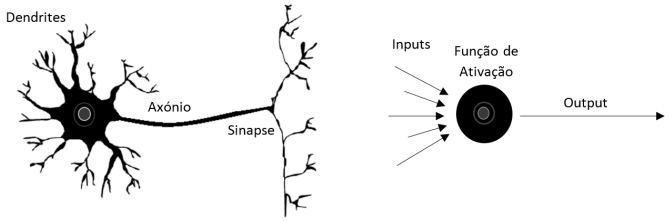


Figura 3.1: Analogia de Comportamento de Neurónios (Biológico e Artificial).

Uma típica rede neuronal é constituída por neurónios artificiais ligados entre si e que são capazes de processar informação. Existe um número elevado de elementos simples chamados unidades e uma rede de ligações direcionadas (com pesos atribuídos) entre eles. Cada unidade processa uma função de um número limitado de saídas de outra unidade da rede, saídas essas que são pesadas e se tornam as entradas da unidade seguinte. Portanto uma

rede não é mais do que um conjunto de nós, em que alguns estão na camada de entrada (onde as unidades recebem os padrões), alguns nas camadas intermédias / escondidas (onde são efetuados o processamento e extração de características) e alguns na camada de saída (que apresenta o resultado final do processamento que se desencadeia na camada intermédia). A lógica da rede neuronal está sobretudo nestes processamentos, que podem ser muito simples (cingindo-se à soma de *inputs*, por exemplo) ou muito complexos (se um nó contiver uma rede neuronal, por exemplo).

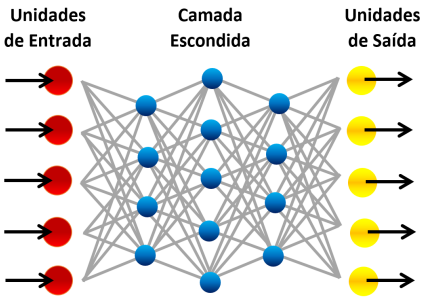


Figura 3.2: Ilustração de uma Rede Neuronal Artificial.

Uma ANN tem a capacidade de aprender com a informação que lhe é fornecida, melhorando o seu desempenho durante o processo de aprendizagem, uma vez que aprende por um processo iterativo de ajuste de pesos (forças sinápticas). Em cada iteração do processo de aprendizagem, apresenta a capacidade de aperfeiçoar a sua representação porque aprende por treino (segundo certas regras pré-definidas). Ao conjunto de regras pré-definidas pelo qual se faz a alteração dos pesos dá-se o nome de algoritmo de aprendizagem, que define a forma como os pesos são corrigidos e qual a estrutura da rede. O algoritmo mais utilizado é o *Backpropagation*.

Por norma, um dos maiores problemas das ANN é a definição da estrutura / arquitetura da rede, isto é, a estimação do número de camadas ocultas e do número de neurónios em cada camada. Cada neurónio recebe impulsos de entrada e calcula a informação de saída como função desses impulsos, pelo que é realizado um cálculo linear inicial nos inputs e, seguidamente, aplicada uma função de ativação (Tabela 3.1).

Função de Ativação	**Expressão**
Linear	$f(x) = x$
Sinusoidal	$f(x) = sin(x)$ $f(x) = cos(x)$
Sigmóide	$f(x) = \frac{2}{1+e^{-x}} - 1$
Gaussiana	$f(x) = e^{-\frac{x^2}{2}}$

Tabela 3.1: Funções de Ativação Usuais na Aplicação de Redes Neuronais Artificiais.

Os sinais resultantes são posteriormente somados e à soma resultante é aplicada uma função não linear - função transferência - que verifica se o valor resultante da soma entre o produto

dos sinais de entrada pelos respetivos pesos atingiu ou não um valor limite pré-definido, sendo assim gerado o *output* [Faraway (2005)].

Uma das arquiteturas de redes mais utilizada é a *feedforward*, também chamada *rede sem realimentação*, que se caracteriza pelo agrupamento de neurónios em camadas e pelo facto de o sinal percorrer a rede numa única direção (da entrada para a saída), não se estabelecendo ligações entre os neurónios de uma mesma camada. Relembre-se ainda que, de acordo com um dos Teoremas da Aproximação Universal [Suykens *et al.* (2012)], "uma rede *feedforward* com **uma única** camada escondida que contém um número finito de neurónios, pode aproximar funções contínuas em subconjuntos compactos de $\mathbb{R}^n$, sob suposições leves na função de ativação" [Gybenko (1989)]. Neste caso, a rede neuronal *feedforward com uma camada escondida* toma a forma

$$f_o(\sum_h w_{h_o} f_h(\sum_i w_{h_i} x_i))$$

onde f_o representa a função de transferência, f_h representa a função de ativação e w representa os pesos das ligações da rede (w_{h_o} nas ligações entre o *input* e os neurónios da camada escondida e w_{h_i} nas ligações entre os neurónios da camada escondida e o *output*). x_i representam os *inputs* da rede. Será esta a rede neuronal utilizada no caso prático, a ver no capítulo de Aplicação.

Observação: Na estimação do número de camadas e neurónios da rede, deverão ser tidos em conta os seguintes dois efeitos possíveis: *underfitting*, caracterizado pela definição de poucos neurónios, não suficientes para que se consiga estabelecer uma rede neuronal fiável (ou para que sejam identificados padrões) e *overfitting*, definida pela existência de muitos neurónios, que são treinados por um número limitado de informação contida no conjunto de dados.

Uma rede neuronal treinada pode ser usada para fornecer **projeções** face a situações de interesse.

3.2.2 Árvores de Regressão

As árvores podem ser usadas em problemas de regressão ou classificação. A principal diferença reside no facto de as folhas das árvores de regressão conterem **previsões numéricas** e não decisões. O objetivo deste método consiste na partição do espaço preditivo do conjunto de treino em regiões, de modo que essas regiões (subconjuntos finais) sejam tão "puras" quanto possível. A partição passo a passo obtida corresponde a uma aproximação da partição ótima.

Para cada nó da árvore, é necessário escolher a variável que melhor segmenta esse nó, definindo-se uma medida de impureza de tal modo que os descendentes do nó sejam mais puros (existindo menos mistura de informação das regiões até aí definidas) do que o nó que lhes deu origem.

No caso de uma árvore de classificação, as **medidas de impureza** para um dado nó que são mais frequentemente utilizadas (por favorecem os nós mais puros quando comparadas com o erro de classificação), são:

- Quantidade de Informação de Shannon: $-\sum_j P(c_j)log_2(P(c_j))$
- Índice de Gini: $1 - \sum_j P^2(c_j)$

Note-se que $P(c_j)$ é a fração das variáveis independentes no nó em análise que pertencem à região c_j (sejam $c = (c_1, c_2, ...)$ as regiões definidas pelo algoritmo de aprendizagem). Estas medidas satisfazem as propriedades definidas em Breiman *et al.* (1984).

No caso de uma árvore de regressão, o custo de escolher o valor $y = a$ num dado nó é em geral determinado por uma das duas seguintes medidas:

$$\mathrm{E}[(Y-a)^2] \text{ ou } \mathrm{E}[|Y-a|].$$

A ação que minimiza o custo, no primeiro caso, é a atribuição a a do valor da média de Y (a =E[Y]), enquanto que no segundo caso é a atribuição a a do valor da mediana de Y ($a = \varepsilon Y$). Por este motivo, as medidas de impureza a considerar em regressão são:

- **Desvio quadrático médio:** $\mathrm{E}[(Y-\mathrm{E}[Y])^2]$
- **Desvio absoluto médio:** $\mathrm{E}[|Y - \varepsilon Y|]$

A medida de impureza utilizada no caso prático será a dada pelo desvio quadrático médio.

3.2.3 Regressão com Máquinas de Suporte Vetorial

A noção de Regressão com Máquinas de Suporte Vectorial (*Support Vector Regression, SVRs*) [Vapnik (1995)] surgiu como uma generalização das Máquinas de Suporte Vetorial (*SVMs*), que por sua vez surgiram para resolver os problemas computacionais do método do núcleo. Estes estavam relacionados com a lentidão de resposta para grandes conjuntos de dados e necessidade de armazenamento de toda a base de dados para fazer previsão. De modo informal, o método do núcleo é um método não paramétrico para estimação de curvas de densidades onde cada observação é ponderada pela distância em relação a um valor central, designado como núcleo, cuja ideia foi introduzida para previsão por Nadaraya (1964) e Watson (1964).

O método SVR é complexo e a sua explicação não será exaustiva, mas intuitiva (ver Figura 3.3). A exploração deste método pode ser consultada, por exemplo, no tutorial Burges (1998). De um modo muito geral, nas SVMs, através da aplicação de uma função de núcleo (K) às observações, estas são projetadas num espaço de maior dimensão no qual os dados podem ser separados por um hiperplano. Quando os dados de treino são separáveis, o **hiperplano ótimo** no espaço característico apresenta a máxima margem de separação.

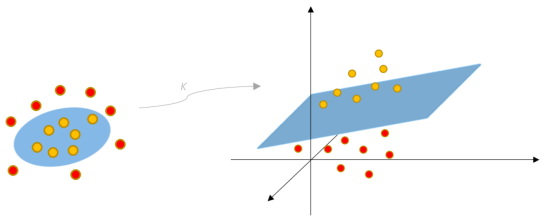


Figura 3.3: Ilustração de uma SVM.

Dado um conjunto de treino $(\mathbf{x}_i, y_i)_i$, onde $\mathbf{x}_i \in \mathbb{R}^n$ e $\mathbf{y} \in \{1, -1\}^l$, procura-se resolver o seguinte problema de otimização:

$$\begin{aligned} \min_{\mathbf{w},\, b,\, \xi} \quad & \frac{1}{2}\mathbf{w}^{\mathbf{T}}\mathbf{w} + C\sum_{i=1}^{l}\xi_i \\ \text{sujeito a} \quad & y_i(\mathbf{w}^{\mathbf{T}}\kappa(\mathbf{x}_i) + b) \geq 1 - \xi_i \quad i = 1, \ldots, l. \\ & \xi_i \geq 0, \end{aligned}$$

onde $\mathbf{x}_i$ representa os vetores de treino, $\mathbf{w}^{\mathbf{T}}$ representa a transposta de $\mathbf{w}$, $C > 0$ o parâmetro de penalização dos erros ξ_i e $K(\mathbf{x}_i, \mathbf{x}_j) \equiv \kappa(\mathbf{x}_i)^{\mathbf{T}}\kappa(\mathbf{x}_j)$ é a chamada função núcleo (ver Tabela 3.2). A função núcleo utilizada será a radial.

Função Núcleo	**Expressão**
Linear	${\mathbf{x}_i}^{\mathbf{T}}\mathbf{x}_j$
Polinomial	$(\gamma {\mathbf{x}_i}^{\mathbf{T}}\mathbf{x}_j + r)^d \quad \gamma > 0$
Radial	$\exp(-\gamma \lvert\lvert {\mathbf{x}_i}^{\mathbf{T}}\mathbf{x}_j \rvert\rvert^2) \quad \gamma > 0$
Sigmóide	$\tanh(\gamma {\mathbf{x}_i}^{\mathbf{T}}\mathbf{x}_j + r)$

Tabela 3.2: Funções Núcleo Usuais na Aplicação de SVMs.

3.3 Avaliação dos Modelos

Para fazer a avaliação dos modelos de previsão, o conjunto de dados é particionado em dois subconjuntos: um de **treino**, usado para a aprendizagem do modelo, e outro de **teste**, para analisar o erro associado ao modelo. Se assim não fosse, o modelo encontrado já estaria ajustado aos dados de teste, produzindo previsões otimistas (e falseando, de algum modo, os resultados).

A separação pode ser realizada por validação cruzada (método *k-fold* [Witten e Frank (2005)]) no caso dos algoritmos que aplicam os métodos de *Data Mining* acima referidos. Porém, é *muito importante* lembrar que, para séries temporais, qualquer forma de reamostragem altera a ordem inicial dos dados e, por isso mesmo, esta técnica não pode ser aplicada, pelo que, nesse caso em particular, foi usado um método designado por *Sliding Window*, que consiste na divisão dos dados existentes em duas janelas: uma que contém as observações

anteriores a um dado instante de tempo e outra que contém as restantes observações, e onde, para a aprendizagem no conjunto de teste, é construído um novo modelo, para cada conjunto de teste, obtido treinando todos os dados anteriores a ele (note-se que, de cada vez que uma nova observação é adicionada ao conjunto de treino, uma mais antiga é removida).

As principais **métricas** usadas para avaliar os erros de previsão dos modelos anteriormente explicados são definidas na Tabela 3.3, onde n representa o número de observações disponíveis, Y_t o valor observado em t e $\hat{Y}_t$ a previsão do modelo para esse mesmo valor.

Denominação	**Acrónimo**	**Expressão**
Erro Quadrático Médio	MSE	$\frac{1}{n}\sum_{i=1}^{n}(Y_t - \hat{Y}_t)^2$
Raíz do Erro Quadrático Médio	RMSE	$\sqrt{\text{MSE}}$
Erro Absoluto Médio	MAD	$\frac{1}{n}\sum_{i=1}^{n}\lvert Y_t - \hat{Y}_t\rvert$
Erro Percentual Absoluto Médio	MAPE	$\frac{100}{n}\sum_{i=1}^{n}\lvert\frac{Y_t - \hat{Y}_t}{Y_t}\rvert$

Tabela 3.3: Métricas Comuns na Avaliação dos Erros dos Modelos de Previsão.

Observe-se que o MSE apresenta uma grande sensibilidade a erros elevados, uma vez que considera o quadrado das diferenças entre os valores observados e previstos. É por esse motivo que muitas vezes se considera o RMSE, que atenua essa desvantagem. O MAD, quando comparado com o MSE, apresenta a vantagem de avaliar os erros na unidade original dos dados (e não ao quadrado), tratando-os de igual modo. O MAPE, que será o erro usado no caso prático, mede o "tamanho do erro" e corresponde à métrica que mais informação dá relativamente à qualidade preditiva do modelo. Note-se que, nesta última métrica, as observações tais que $Y_t = 0$ não podem ser avaliadas.

Uma vez definidas as ferramentas teóricas fundamentais propostas para utilização, é possível continuar a análise, agora sobre a aplicação das mesmas a dados reais.

Capítulo 4

Aplicação

A base de dados analisada apresenta algumas variáveis climáticas (referidas na Tabela 4.1 e cuja nomenclatura será, daqui em diante, a presente na segunda coluna da mesma) cujas características são detalhadas no Anexo D.

Variável	**Abreviatura**	**Unidade de Medida**
Ano	A	-
Mês	M	-
Dia	D	-
Hora	H	-
Estação do Ano	E	-
Direção do Vento Horária	DV	Graus
Precipitação Total Horária	PT	mm
Temperatura Mínima Horária	TMin	Graus Celsius
Temperatura Média Horária	TMed	Graus Celsius
Temperatura Máxima Horária	TMax	Graus Celsius
Humidade Relativa Mínima Horária	HMin	%
Humidade Relativa Média Horária	HMed	%
Humidade Relativa Máxima Horária	HMax	%
Radiação Global Mínima Horária	RMin	W/m^2
Radiação Global Média Horária	RMed	W/m^2
Radiação Global Máxima Horária	RMax	W/m^2
Velocidade de Vento Mínima Horária	VMin	m/s
Velocidade de Vento Média Horária	VMed	m/s
Velocidade de Vento Máxima Horária	VMax	m/s

Tabela 4.1: Variáveis da Base de Dados.

Para se avaliar quais as variáveis com maior influência sobre a velocidade máxima de vento, foi efetuada uma pequena análise, precedida de um tratamento da base de dados. Neste último, executou-se :

- preenchimento de valores em falha (vulgarmente designados por NA - *Not Available*), que surgiam em seis das variáveis (HMed, HMin e HMax, em aproximadamente 3%

do tamanho da amostra; PT, em duas observações; VMin e VMax, numa observação), substituindo-os pelo correspondente valor da mediana da variável (seguindo a abordagem sugerida por Torgo (2010)).

- dada a eventual importância que a variável temporal "Data" poderá ter no estudo, esta foi substituída por cinco outras variáveis: Ano, Mês, Dia, Hora e Estação do Ano.

Tal como escrito anteriormente, o foco incidirá sobre a variável *vento*, em particular sobre a variável **velocidade máxima do vento** da base de dados.

Como as instruções implementadas em R para o cálculo da correlação não aceitam variáveis categóricas, a variável Estação do Ano foi tranformada em numérica apenas para que este cálculo pudesse ter em conta todas as variáveis da base de dados. Note-se que, dado um conjunto de dados com diversas variáveis, se o objetivo for diminuir o seu número, não se deve considerar as que estão altamente correlacionadas entre si (uma vez que contêm a mesma informação, tornando-se redundantes). Pela observação de boxplots e pela matriz de correlação entre variáveis, não aparenta existir uma dependência clara entre a variável em estudo e as restantes (exceto no caso em que a variável climática é a mesma, mas analisada em relação aos seus valores máximo, mínimo e médio horários).

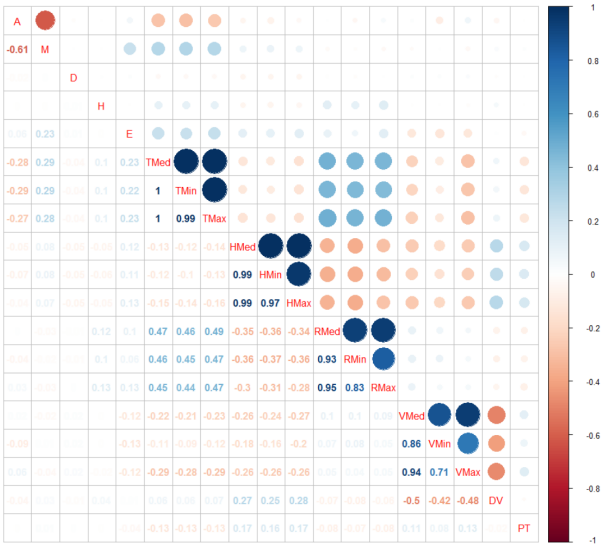


Figura 4.1: Correlações Amostrais.

No entanto, o facto de não aparentar existir correlação linear entre as variáveis, quando analisadas duas a duas, não significa que a variável em estudo não possa ser descrita por uma combinação das outras variáveis. Nesse caso, poder-se-ia proceder ao estudo dessa relação considerando a velocidade máxima de vento como variável resposta de um **modelo de regressão linear múltipla** e essa variável, seguidamente designada por V, poderia ser escrita como

$$V = \beta_0 + X\beta_1 + ... + \beta_p X_p + u = X\beta$$

onde $X_1, ..., X_p$ são as variáveis explicativas, u os erros (também designados por resíduos) do modelo e $\beta = (\beta_0, \beta_1, ..., \beta_p)$ os coeficientes da regressão a estimar.

No modelo clássico de regressão linear assume-se que os erros u são i.i.d., $u \sim N(0; \varepsilon = \sigma^2 Id)$ e V segue uma distribuição normal com média dependente de modo linear das variáveis explicativas, isto é, $V|X \sim N(\mu(X), \sigma^2(X)), \quad E(V|X) = \beta_0 + X\beta_1 + ... + \beta_p X_p$.

De um modo sucinto, face à expressão geral de definição do modelo de regressão linear e para que se faça a correta interpretação dos coeficientes estimados no modelo, é importante referir que:

- O coeficiente independente, β_0, representa a resposta caso todas as variáveis explicativas sejam nulas.
- Para $j = 1, ..., p$ o coeficiente j representa o incremento médio de V quando a variável explicativa X_j é aumentada de uma unidade e as restantes variáveis explicativas se mantêm constantes, o que permite avaliar a intensidade da relação entre V e X_j.

A primeira avaliação dos modelos deve ser feita por **análise** de **testes de hipóteses**, **critérios de informação** e **resíduos**. Antes de se observarem os resultados, é necessário salientar alguns aspetos:

- Como o problema a resolver é o da estimação dos parâmetros β e σ^2, mantendo σ^2 fixo, β pode ser estimado pelo método da máxima verosimilhança, pelo que, supondo a independência entre as observações, a função de verosimilhança a maximizar será dada por

$$L(\beta, \sigma^2 \mid (v_i, x_i)_i) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\Big(-\frac{1}{2} \frac{(v_i - \hat{v}_i)^2}{\sigma^2} \Big).$$

  Como a função logaritmo é crescente, maximizar L equivale a maximizar $log(L)$, ou seja

$$l(\beta, \sigma^2 \mid (v_i, x_i)_i) = log(L(\beta, \sigma^2 \mid (v_i, x_i)_i))$$

$$= -\frac{n}{2} log(2\pi\sigma^2) - \frac{1}{2} \sum_{i=1}^{n} \frac{(v_i - \hat{v}_i)^2}{\sigma^2} = -\frac{n}{2} log(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (v_i - X_i.\beta)^2.$$

  Os critérios de informação são usados para comparação de modelos não encaixados e aplicados a modelos construídos a partir da maximização do logaritmo da verosimilhança acima definido. Caracterizam-se pela penalização de modelos com maior número de parâmetros e os mais usados são, segundo Pinheiro e Bates (2000):

  - critério de informação de **Akaike** (AIC), dado por $-2l(\hat{\beta}, \hat{\sigma}^2) + 2n_{\text{par}}$
  - critério de informação **Bayesiana** (BIC), dado por $-2l(\hat{\beta}, \hat{\sigma}^2) + 2n_{\text{par}} log(N)$

  onde n_{par} é o número de parâmetros do modelo e N o número de observações.

- Os principais testes envolvidos no teste de um modelo de regressão linear são os seguintes [Faraway (2009)]:
  - teste de **Wald**, que indica a significância para o modelo de cada um dos coeficientes estimados, testando a hipótese nula de que $\beta_j = 0$ para algum $j = 0, 1, ..., p$; a análise dos resultados deste teste deve ser feita com precaução, uma vez que a rejeição da hipótese nula não implica que todas as variáveis não significativas devam ser excluídas do modelo. A remoção de uma variável explicativa pode fazer com que outras, que antes não eram consideradas significativas, o passem a ser.
  - teste-**F**, através do qual se pretende averiguar se um determinado grupo de variáveis explicativas é ou não significativo na explicação da variável resposta do modelo de regressão, testando a hipótese nula de que $\beta_0 = \beta_1 = ... = \beta_p = 0$, e a alternativa dada por $\exists\, j \in \{1, ..., p\}$ tal que $\beta_j \neq 0$. É usual apresentar o resultado de uma análise de regressão sob a forma de uma tabela de análise da variância, ANOVA, onde se indicam alguns valores necessários ao desenvolvimento do teste de hipóteses anterior.
    Além disso, a variação da variável de resposta é decomposta na soma da variação devida à regressão (*Regression Sum of Squares*) e da variação residual (*Residual Sum of Squares*). Por esse motivo, tal como intuitivamente faz sentido, um modelo que apresente um bom ajustamento aos dados será um modelo em que a variação total será essencialmente devida à regressão, apresentando uma variação residual baixa.

Neste âmbito, foram avaliados modelos com e sem algumas das variáveis. As referentes aos valores médios, uma vez que correspondem a valores calculados e não medidos, não foram consideradas. Como visto na Figura 4.1, a informação recolhida mostrou que algumas das variáveis são altamente correlacionadas. Para decidir quais dessas variáveis se deveria incluir no modelo, optou-se pelas menos correlacionadas entre si. Consequentemente, retiveram-se as variáveis TMin, HMax e RMin em detrimento das variáveis TMax, HMin e RMax, respetivamente.

As diferentes aproximações do modelo por regressão linear sugeriram que algumas das variáveis não eram significativas (valor-p superior a 0.05). Estas foram removidas uma a uma e os resultados analisados. O resultado final dessa análise, obtido em R, foi o seguinte:

```
Call:
lm(formula = VMax ~ I(A) + I(M) + I(Estacao) + TMin + HMax +
RMin + DV + PT, data = BaseDados)

Residuals:
Min       1Q   Median       3Q      Max
-16.0302  -1.9165  -0.2585   1.6601  18.3484

Coefficients:          Estimate  Std. Error  t value Pr(>|t|)
(Intercept)         -1.941e+02  9.442e+01  -2.056   0.0398 *
I(A)                 1.025e-01  4.686e-02   2.188   0.0287 *
I(M)                 2.951e-02  1.328e-02   2.222   0.0263 *
I(Estacao)Outono     4.924e-01  1.242e-01   3.966 7.35e-05 ***
```

```
I(Estacao)Primavera -8.014e-01  8.810e-02  -9.097  < 2e-16 ***
I(Estacao)Verão     -3.265e-01  1.288e-01  -2.535   0.0112 *
TMin                -1.795e-01  7.098e-03 -25.294  < 2e-16 ***
HMax                -3.998e-02  1.519e-03 -26.330  < 2e-16 ***
RMin                 2.204e-03  1.537e-04  14.341  < 2e-16 ***
DV                  -1.347e-02  2.179e-04 -61.822  < 2e-16 ***
PT                   4.609e-01  2.431e-02  18.955  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.227 on 17732 degrees of freedom
Multiple R-squared:  0.3568,Adjusted R-squared:  0.3564
F-statistic: 983.4 on 10 and 17732 DF,  p-value: < 2.2e-16
```

Comentário: A rejeição da hipótese nula no teste-F não significa que o modelo seja bom. Aliás, observando o valor do coeficiente de determinação (quadrado do coeficiente de correlação linear de Pearson amostral entre os valores observados e os ajustados), que deveria ser próximo de 1, pode concluir-se que o modelo apresenta uma variação residual alta (o coeficiente de determinação é igual a 0.3568), pelo que é um indicador de um **mau ajustamento** aos dados. A inexistência de relação linear entre a variável resposta e as variáveis explicativas conduz a valores de coeficiente de determinação próximos de 0. Este também não parece ser o caso. Vai ser necessário analisar os resíduos e outras medidas de significância.

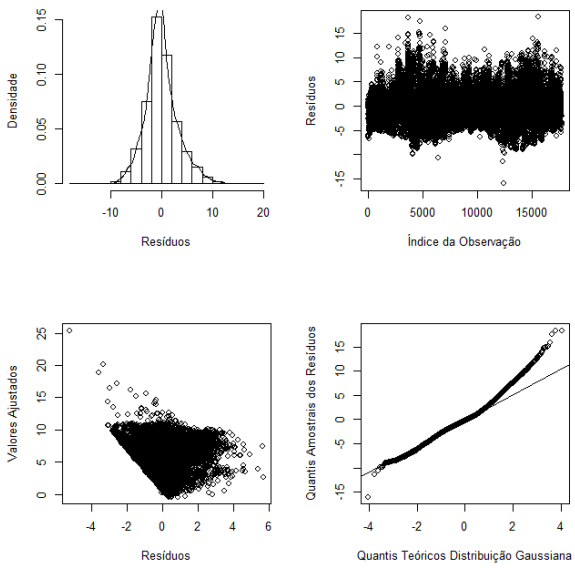


Figura 4.2: Análise Gráfica de Resíduos (caso horário).

Chegou-se à conclusão que o modelo apresenta 2485 observações com resíduo superior a 3.3 que, segundo a literatura da área, é o valor de referência a partir do qual as observações devem ser analisadas uma a uma, com o intuito de perceber qual a sua influência sobre o modelo estimado pela regressão, o que não parece ser exequível dado o elevado número de observações nessas condições!

Com o mesmo objetivo, analisaram-se as distâncias de Cook [1] e *leverages* (ou pontos de alta influência) obtidos pelo modelo. Pontos com distância de Cook superior a $\frac{4}{N}$ (mantendo a notação anterior) podem ser pontos influentes pelo que o seu impacto no ajustamento do modelo deve ser averiguado e, nesse caso, quase toda a amostra deveria ser analisada. Até agora, tudo aponta para que esta **não seja uma boa solução de ajuste** à variável resposta V. Como último elemento de análise, foi estudado o fator de inflação da variância para X_i (VIF [2]), que é uma medida de aumento da variância de β_i quando as restantes variáveis explicativas estão correlacionadas. Este foi o principal critério para eliminar modelos que apresentavam valores superiores de coeficiente de determinação, mas que também apresentavam muticolinearidade (i.e, VIF>10 para algum X_i). Note-se que a multicolinearidade pode interferir com os parâmetros estimados, falseando a resposta dada pelo modelo de regressão [Belsley *et al.* (1980)].

A análise gráfica do histograma mostra o que não parece ser uma má aproximação de uma distribuição normal, mas a análise do gráfico dos quantis amostrais contra os quantis de uma distribuição gaussiana mostra que, nos valores de cauda, os resíduos se afastam do comportamento esperado para uma normal. Finalmente, fez-se a análise dos resíduos estandardizados contra os valores ajustados (gráfico inferior esquerdo da Figura 4.2). Deveria observar-se a variância constante na direção dos resíduos e as observações deveriam formar uma nuvem sem tendência, mas não é esse o padrão observado. A inexistência deste padrão sugere **heterocedasticidade e / ou não linearidade**, o que, pela análise do gráfico anterior, sugere que seja esta a situação dos dados analisados.

Assumiu-se, até agora, que a matriz de covariância entre os erros, ε, é tal que $\varepsilon = \sigma^2 Id$, mas pode acontecer que os resíduos tenham variância não constante ou estejam correlacionados. Para testar esta última hipótese, usou-se o teste de Durbin e Watson (1950), que **rejeitou** a existência de **independência** entre os resíduos. Considere-se, por esse motivo, um outro modelo de regressão normalmente usado em estudos de dados longitudinais (caracterizados pela dependência das observações ao longo do tempo) e cujo detalhe teórico e prático pode ser consultado em Cabral e Gonçalves (2011). Suponha-se, então, que $\varepsilon = \sigma^2\Sigma$, onde Σ é conhecido e σ^2 é desconhecido. Assim, o **método dos mínimos quadrados generalizados** (GLS - *General Least Squares*) minimiza (mantendo a notação anterior)

$$(v - X\beta)^{\mathrm{T}}\Sigma^{-1}(v - X\beta)$$

resolvendo

$$\hat{\beta} = (X^{\mathrm{T}}\Sigma X)^{-1}X^{\mathrm{T}}\Sigma^{-1}v.$$

Escrevendo $\Sigma = SS^{\mathrm{T}}$, onde S é uma matriz triangular, $v' = S^{-1}v$, $X' = S^{-1}X$ e $\varepsilon' = S^{-1}\varepsilon$, tem-se, pela decomposição de Cholesky [Brezinski e Zaglia (2005)], que

$$v = X\beta + \varepsilon \implies S^{-1}v = S^{-1}X\beta + S^{-1}\varepsilon \implies v' = X'\beta\varepsilon'.$$

[1] dadas por $\frac{\sum_{j=1}^{n}(\hat{v}_j - \hat{v}^2_{j/i}}{1}$, onde $\hat{v}^2_{j/i}$ representa o valor ajustado para v_j quando o modelo é estimado excluindo a observação i.

[2] dado por $\mathrm{VIF}(X_i) = \frac{1}{1-{R_i}^2}$ onde ${R_i}^2$ é o coeficiente de determinação de X_i sobre as restantes variáveis explicativas.

Analisando a variância dos resíduos da nova regressão considerada, tem-se que

$$\text{var } \varepsilon' = \text{var}(S^{-1}\varepsilon) = S^{-1}(\text{var } \varepsilon)S^{\text{-T}} = S^{-1}\sigma^2 S S^{\text{T}} S^{\text{-T}} = \sigma^2 Id$$

portanto as variáveis v' e X' estão relacionadas por uma equação de regressão com erros não correlacionados com igual variância. No entanto, existem funções de variância que podem ser usadas para modelar a estrutura de variância dos erros e que são apresentadas na Tabela 4.2, onde ϱ_{it} é o vetor de covariáveis tais que $i = 1, ..., n$; $t = 1, ..., T_i$, δ é o vetor de parâmetros da variância, s_{it} uma variável de estratificação e δ_1, δ_2 parâmetros das funções de variância.

Classe	**Variância**
VarFixed - variância com covariável única	$\sigma^2 \varrho_{it}$
VarIdent - variância que difere por categoria da covariável	$\sigma^2 \delta^2_{s_{it}}$
VarPower - potência de uma covariável	$\sigma^2 \lvert\varrho_{it}\rvert^{2\delta}$
VarExp - exponencial de uma covariável	$\sigma^2 \text{exp}(2\delta\varrho_{it})$
VarConstPower - constante + potência de uma covariável	$\sigma^2(\delta_1 + \lvert\varrho_{it}\rvert^{\delta_2})^2$

Tabela 4.2: Funções de Variância para a Modelação da Heterocedasticidade.

Comentário: Aplicando o método dos mínimos quadrados generalizados ao caso de estudo, os resultados da significância dos parâmetros estimados assemelharam-se aos obtidos na regressão linear (o resultado pode ser visto no Anexo A). No entanto, esta revelou-se mais eficaz no tratamento da variância. Concluiu-se, pelos gráficos dos resíduos estandardizados em cada uma das variáveis do modelo, que os resíduos estandardizados associados às variáveis DV e TMin apresentavam tendência. Foram usadas diferentes funções de variância para modelar a variância dessas variáveis e foi feita a respetiva avaliação do modelo GLS produzido (por avaliação de critérios de informação, uma vez que algumas das combinações testadas obrigavam à avaliação de dois modelos não encaixados - isto é, em que os parâmetros de um dos modelos não constituíam um subconjunto dos parâmetros do outro modelo). Foram testadas:

- função VarPower aplicada variável TMin.
- função VarIdent aplicada à variável E, uma vez que foi detetada uma pequena diferença visual nos resíduos por Estação do Ano.
- função VarPower aplicada à variável TMin, restrita por E. Deste modo, vai ser estimado um parâmetro por estação para a covariável TMin.
- foram testados os três modelos anteriores, substituindo a função VarPower pela função VarExp.

Comentário: Verificou-se que o melhor modelo foi o que usou a função VarExp sobre a covariável TMin, definindo um parâmetro por Estação do Ano. Isto significa que o modelo estimou uma variância diferente por estação que diminui (uma vez que os parâmetros estimados foram todos negativos) de forma exponencial com a Temperatura Mínima. Apresentou valores de AIC=88984.99 e BIC=89109.52, os mais baixos de todos os testados. Os resultados podem ser vistos no Anexo A.

A rejeição da independência dos erros na regressão também pode ser um fator importante para explicar a alta componente residual nos modelos vistos até agora. Por esse motivo, consideraram-se os valores máximos **diários** de vento (escolhendo o valor máximo horário de cada dia) e aplicou-se regressão linear múltipla nos mesmos moldes que anteriormente. A análise gráfica de resíduos, nesse caso, é apresentada na Figura 4.3:

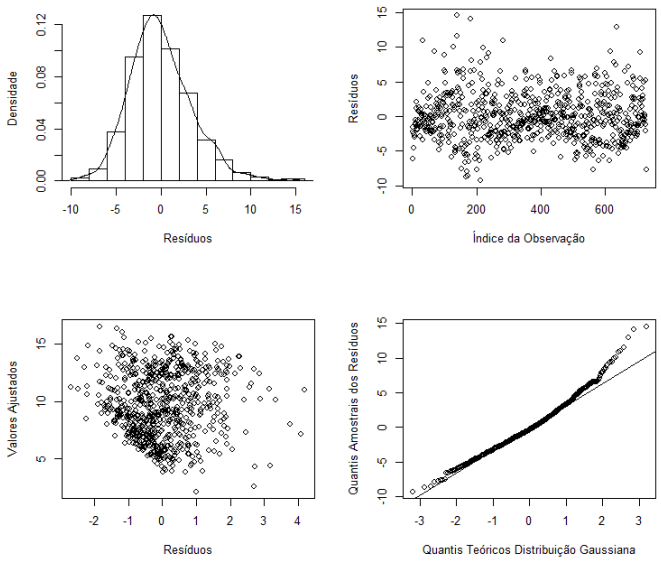


Figura 4.3: Análise Gráfica de Resíduos (caso diário).

Apesar dos testes de independência entre resíduos continuarem a rejeitar a hipótese nula, a variância dos resíduos já não parece comportar tanta informação como no caso visto anteriormente, o que intuitivamente faz sentido, visto que se passou de 17743 observações para 730.

Olhando agora para o histograma da principal variável em estudo, é possível observar que a forma da mesma não aparenta seguir distribuições conhecidas. Assim, face à dificuldade em modelar a distribuição seguida pelas velocidades de ventos desta estação meteorológica (após diversas tentativas de ajuste a várias distribuições, com estimação dos respectivos parâmetros por método de máxima verosimilhança, e execuções de testes de hipóteses tipo Kolmogorov-Smirnov para comparação de duas distribuições - a empírica com a estimada - com valores-p sempre inferiores ao valor de referência 0.05) optou-se por uma estimação não paramétrica da função densidade de probabilidade da velocidade máxima dos ventos, dada por uma modificação do método do núcleo e explicada em Bessa *et al.* (2012), cujo resultado se apresenta na Figura 4.4.

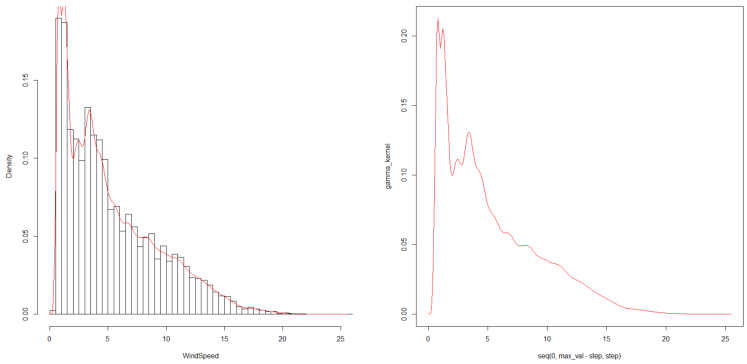


Figura 4.4: Função Densidade de probabilidade Estimada pelo Método do Núcleo Modificado.

Esta informação é mostrada apenas para corroborar os resultados vistos. Efetivamente, a variável resposta dos modelos usados até agora não tem um comportamento que possa ser associado à distribuição normal. Também foram feitas transformações dos dados (raíz quadrada, logaritmo e transformação de Box) e feita uma análise semelhante à referida, mas os resultados não melhoraram, pelo que se optou pela não colocação desse trabalho na tese. A tentativa de ajuste da função densidade de probabilidade dos dados às distribuições que caracterizam tipicamente este tipo de fenómeno, como Weibull, Gumbel ou GEV (*Generalized Extreme Value*) também não se mostrou eficaz.

4.1 Modelação

Relativamente ao processo de **modelação de velocidade de ventos**, e antes da apresentação de resultados, deve explicar-se alguns detalhes computacionais.

Detalhes de Implementação:

- É comum não se considerarem as transições de um estado para si mesmo. Nesse sentido, na implementação em R são consideradas apenas as transições de um estado i para um estado j, com $i \neq j$ (logo $p_{ii} = 0$). Para o processo de simulação, assumiu-se que $p_{ij}(s) = p_{ij}(0) \quad \forall s$.

- Para estimar $F_{i,j}$, é definida uma matriz com três dimensões: número de linhas (estado para onde o sistema transitou na $(n-1)$-ésima transição), número de colunas (estado para onde o sistema transitou na n-ésima transição) e tempos considerados para estudo. Isto acontece porque, na verdade, vão existir tantas matrizes F quantas o número de unidades temporais estudadas. Para a definição desta função de distribuição, foi definida uma variável auxiliar temporária *ftemp* que guarda o número de vezes que o sistema se encontra no estado i e transita para o estado j ao fim de t unidades de tempo. Assim, o objetivo desssa variável auxiliar é fazer a contagem das unidades de tempo em que o sistema permanece num dado estado. Para exemplificar, se o sistema em estudo fosse o seguinte, 11111222111112, então $ftemp[1,2,5] = 2$ e $ftemp[2,1,3] = 1$. Uma vez que *ftemp* contém a contagem das vezes que o sistema esteve no estado i e transitou para j ao fim de t unidade de tempo, é possível calcular agora a função de probabilidade F, bastando para isso dividir esse número de ocorrências pelo número total de vezes que se transitou de i para j.

- É ainda necessário referir que o modelo de segunda ordem considerado para as simulações não é exatamente como o definido na contextualização teórica, na medida em que é de segunda ordem nos estados mas é de primeira ordem no tempo. Difere do de primeira ordem apenas na consideração não só do estado onde o sistema se encontra, mas também do estado de onde o sistema saiu.

Algoritmo:

Simulação de uma cadeia semi-markoviana até $t = T_{\max}$:

1. Definir $T_0 = 0$, $n = 0$.

2. Definir J_0 como o estado do sistema original.

3. Enquanto $t < T_{\max}$: GERAR J_n a partir de $P_{J_n,*}(T_{n-1})$; Definir $n = n + 1$; GERAR F; Atribuir $t = t + F; t_n = t$
   Caso contrário: pára o processo.

4. Voltar ao primeiro passo.

Usando este algoritmo, foi possível simular dados sintéticos de velocidade de vento.

Observações:

- Note-se que nas simulações computacionais se utiliza sempre a probabilidade acumulada, uma vez que com a função densidade de probabilidade não se conseguem distinguir as probabilidades de cada elemento. Quando se gera uma determinada função de probabilidade, é comum fazê-lo pela geração de um número aleatório entre 0 e 1, para que se possa fazer pleno uso da definição de probabilidade.

- Para não criar uma quarta dimensão e não se ter de alterar o código definido para as simulações markovianas e semi-markovianas de primeira ordem, foi decidido que, para a simulação semi-markoviana de segunda ordem (onde, recorde-se, existem três estados de referência para o sistema: onde estava, onde está e para onde vai após t unidades de tempo), se codificariam os estados de 1 a 36, correspondência associada a todas as combinações de onde estava o sistema e onde está agora. Assim, o algoritmo usado foi o visto anteriormente, uma vez que a informação a considerar já está incluída na codificação dos estados. Essa codificação é mostrada na Tabela 4.3.

  É necessário referir que nem todas as transições são possíveis. Exemplificando, a cadeia 2 1 significa que o sistema estava em 1, está em 2 e vai para 1 após t unidades de tempo. A cadeia 11 25 significa que o sistema estava em 2, está em 5 e vai para 1 após t unidades de tempo. O estado intermédio tem de ser comum entre a primeira transição analisada e a segunda, para que esta exista efetivamente. Por essa razão, a cadeia 4 7, por exemplo, não faz sentido. Este modo de entrada dos estados é usado apenas para a simulação e é realizada a transformação inversa no final do processo para que se tenha o mesmo número de estados nos resultados finais (e, consequentemente, para que seja possível a sua análise e comparação).

Transição	Estado	Transição	Estado	Transição	Estado
$1 \to 1$	1	$2 \to 1$	7	$3 \to 1$	13
$1 \to 2$	2	$2 \to 2$	8	$3 \to 2$	14
$1 \to 3$	3	$2 \to 3$	9	$3 \to 3$	15
$1 \to 4$	4	$2 \to 4$	10	$3 \to 4$	16
$1 \to 5$	5	$2 \to 5$	11	$3 \to 5$	17
$1 \to 6$	6	$2 \to 6$	12	$3 \to 6$	18
Transição	Estado	Transição	Estado	Transição	Estado
$4 \to 1$	19	$5 \to 1$	25	$6 \to 1$	31
$4 \to 2$	20	$5 \to 2$	26	$6 \to 2$	32
$4 \to 3$	21	$5 \to 3$	27	$6 \to 3$	33
$4 \to 4$	22	$5 \to 4$	28	$6 \to 4$	34
$4 \to 5$	23	$5 \to 5$	29	$6 \to 5$	35
$4 \to 6$	24	$5 \to 6$	30	$6 \to 6$	36

Tabela 4.3: Correspondência de Estados usada na Simulação Semi-markoviana de Segunda Ordem.

Com base em todo o processo explicado na secção teórica de modelação desta tese, foi obtida a matriz de transições do sistema (ver equação (2.1) referida na Introdução aos Processos Semi-Markovianos). A exploração da mesma poderia ser numérica, mas optou-se pela análise gráfica, uma vez que permite uma leitura mais intuitiva acerca do potencial da informação obtida por este modelo. A variável estudada foi discretizada em 6 estados: dois de vento fraco, dois de vento intermédio e dois de vento forte (ver Tabela 3 do Anexo C). Alguns dos gráficos obtidos pelo estudo de $\phi_{ij}(1,t), t = 1, ..., T_{\max}$ (efetuado a partir da primeira hora registada na base de dados, para $T_{\max} = 50$ horas) foram os seguintes:

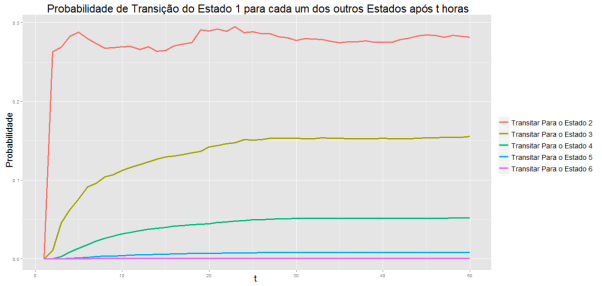
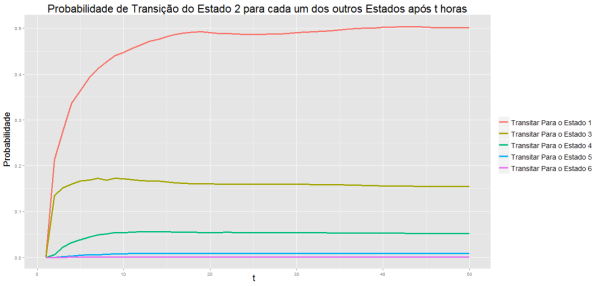
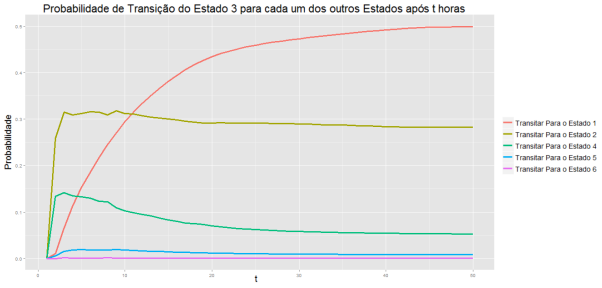
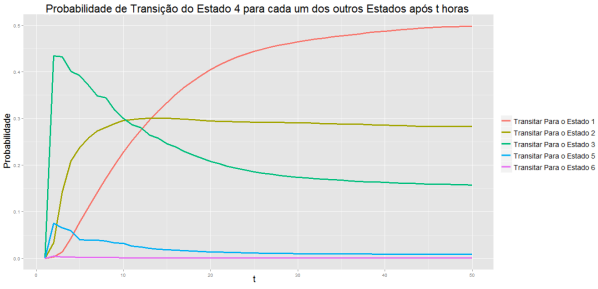
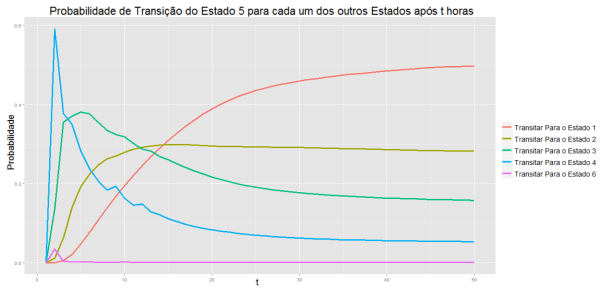
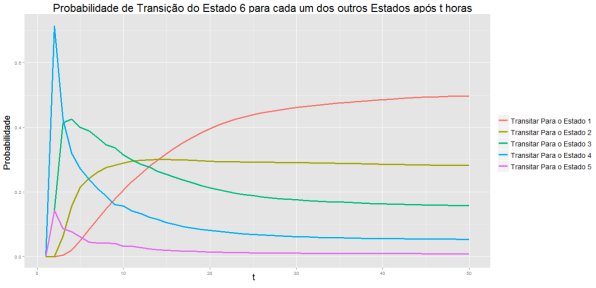
Para $i = 1$	Para $i = 2$
	
Para $i = 3$	Para $i = 4$
	
Para $i = 5$	Para $i = 6$
	

Tabela 4.4: Probabilidades de Transição Estimadas (da Cadeia Semi-markoviana de Primeira Ordem).

Comentário: A análise gráfica permite constatar que, no final das 50 horas analisadas, existe uma estabilização dos valores das probabilidades de transição. É também possível observar que, nos estados de **vento fraco** (1 e 2), a probabilidade de transição para estado de vento forte são muito baixas e que o acontecimento mais provável é a transição para o

outro estado de vento fraco. Quanto aos estados de **vento forte** (5 e 6), é interessante verificar que, após algumas horas, é mais provável que o sistema retorne a um estado de vento fraco do que se mantenha num de vento forte ou **intermédio** (3 e 4). Este facto permite, de algum modo, confirmar o que o senso comum nos diz: "depois da tempestade vem a bonança". No caso do estado 5, nas primeiras horas, é mais provável que transite para um estado intermédio (3 ou 4), ao passo que no estado 6 é mais provável que transite para um estado de vento forte (5). As primeiras 15 horas também parecem ser determinantes no decorrer do processo, em particular nos estados de **vento intermédios** e nos estados de **vento forte**, na medida em que a probabilidade de transição para estados com velocidade de vento próximas das que o estado original apresenta são mais prováveis nessas primeiras horas do que nas seguintes. Note-se que só se observou o comportamento do sistema nas primeiras horas uma vez que a matriz de transição P considerada é igual para todos os instantes de tempo. Se fosse diferente para cada instante, o estudo de outros domínios temporais poderia ser mais informativo.

Esta função fornece quase toda a informação necessária sobre o processo, uma vez que é possível ao leitor estudar as probabilidades de transição entre os estados de interesse, nos instantes de interesse. A **exploração desta informação** seria útil, por exemplo, caso existissem várias regiões em análise. Nesse caso, seria possível a uma qualquer companhia seguradora analisar qual o comportamento esperado do vento nas diferentes regiões, fazendo diferentes atribuições de prémios (*pricing*) consoante a região do segurado. Do mesmo modo, esta informação poderia ser usada para estudar o potencial energético de uma determinada região (acrescentando ao estudo outro tipo de variáveis), aliando informação relativa não só à velocidade de vento, mas também à sua direção [D'Amico *et al.* (2012)].

Fez-se uma análise de diagnóstico com base num histograma para se perceber se este seria semelhante à série original de forma a testar empiricamente se o cálculo da matriz F foi feito corretamente e se as cadeias semi-markovianas (de primeira e segunda ordens) conseguiam modelar os dados ou não. Os resultados obtidos relativos às simulações foram os seguintes:

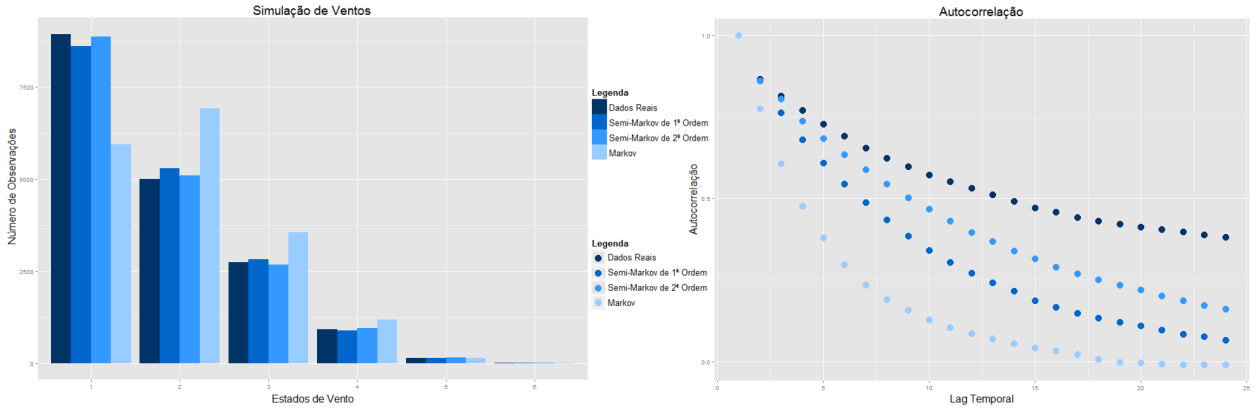


Figura 4.5: Simulações e respetivas Funções de Correlação.

Comentário: Em qualquer um dos gráficos, os processos semi-markovianos aparentam fazer uma modelação que se aproxima do processo original de um modo mais coerente do que o processo markoviano. Note-se que o gráfico da esquerda apenas permite avaliar a

representatividade dos estados no final do processo, mostrando o número de vezes que o sistema esteve em cada um dos estados. A observação do gráfico de autocorrelação (lado direito da Figura 4.5) também corrobora a primeira afirmação, na medida em que a análise temporal definida por esta medida sugere um comportamento por parte dos processos semi-markovianos mais próximo do original, quando comparada com a autocorrelação do processo markoviano.

No entanto, a análise gráfica não basta para fazer conclusões. Para resultados mais elucidativos, mostram-se tabelas com quantidades de interesse, resultantes de 1000 simulações Monte Carlo dos três processos. A Tabela 4.5 exibe as probabilidades médias de transição estimadas (medidas em percentagem e arredondadas a duas casas decimais) e a Tabela 4.6 os tempos médios de espera estimados (horários, arredondados a duas casas decimais) antes de ocorrer transição. Os resultados mais próximos dos originais (avaliados pela diferença em valor absoluto) estão assinalados a azul [3].

Reais

	1	2	3	4	5	6
1	–	96.28	3.72	0.00	0.00	0.00
2	64.12	–	34.67	1.21	0.00	0.00
3	1.34	60.22	–	37.10	1.12	0.22
4	0.23	4.22	77.00	–	18.08	0.47
5	0.00	1.14	14.77	80.68	–	3.41
6	0.00	0.00	14.28	71.43	14.29	–

Markov

	1	2	3	4	5	6
1	–	95.64	4.36	0.00	0.00	0.00
2	62.46	–	35.89	1.65	0.00	0.00
3	1.91	63.74	–	33.06	1.29	0.00
4	0.16	6.13	78.39	–	14.84	0.48
5	0.00	0.00	16.97	78.57	–	4.46
6	0.00	0.00	12.50	75.00	12.50	–

Semi-Markov de Ordem 1

	1	2	3	4	5	6
1	–	95.86	4.14	0.00	0.00	0.00
2	65.59	–	33.10	1.31	0.00	0.00
3	1.24	57.80	–	39.28	1.57	0.11
4	0.44	2.44	78.93	–	17.52	0.67
5	0.00	1.08	15.05	79.57	–	4.30
6	0.00	0.00	0.00	100.00	0.00	–

Semi-Markov de Ordem 2

	1	2	3	4	5	6
1	–	96.40	3.60	0.00	0.00	0.00
2	65.52	–	33.44	1.04	0.00	0.00
3	1.48	60.27	–	36.77	1.25	0.23
4	0.25	5.42	78.08	–	15.76	0.49
5	0.00	2.59	12.99	84.42	–	0.00
6	0.00	0.00	25.00	25.00	50.00	–

Tabela 4.5: Probabilidades Médias de Transição Estimadas (em %).

Comentário: A informação presente na primeira tabela **não indica** que exista efetivamente um melhor ajuste dos modelos semi-markovianos. Apesar de grande parte dos melhores resultados estar assinalada sobre os resultados da modelação semi-markoviana, ainda há situações em que o modelo markoviano é o melhor. De facto, existem situações em que os processos semi-makovianos se aproximam dos originais (como por exemplo nas transições $1 \to 2$, $2 \to 4$, $4 \to 3$ e $5 \to 6$) em detrimento do de Markov. Há outras situações em que esse facto não se verifica (como por exemplo nas transições do estado 6 para qualquer um dos outros estados). Será necessário um critério de decisão mais forte para decidir qual o processo que melhor aproxima as características do processo original.

[3] caso existam valores iguais, são todos assinalados. Para esta avaliação são consideradas apenas as probabilidades e tempos diferentes de 0.

Por esse motivo, foram analisados os tempos de espera, onde os modelos semi-markovianos apresentam claramente melhores resultados do que o modelo markoviano, em particular nos estados de vento fraco e intermédio.

Reais						
	1	2	3	4	5	6
1	–	8.56	9.44	0.00	0.00	0.00
2	3.42	–	3.10	4.06	0.00	0.00
3	1.33	3.03	–	3.21	1.60	4.50
4	1.00	1.28	2.12	–	2.56	1.00
5	0.00	1.00	1.08	1.80	–	1.00
6	0.00	0.00	1.00	1.00	1.00	–

Markov						
	1	2	3	4	5	6
1	–	3.66	3.65	0.00	0.00	0.00
2	2.83	–	2.83	2.80	0.00	0.00
3	2.44	2.45	–	2.45	2.43	0.00
4	1.82	1.81	1.82	–	1.83	1.83
5	0.00	1.30	1.30	1.29	–	1.30
6	0.00	0.00	1.00	1.00	1.00	–

Semi-Markov de Ordem 1						
	1	2	3	4	5	6
1	–	8.55	9.47	0.00	0.00	0.00
2	3.43	–	3.10	4.07	0.00	0.00
3	1.33	3.03	–	3.20	1.60	4.42
4	1.00	1.28	2.12	–	2.57	1.00
5	0.00	1.00	1.08	1.81	–	1.00
6	0.00	0.00	1.00	1.00	1.00	–

Semi-Markov de Ordem 2						
	1	2	3	4	5	6
1	–	8.55	9.43	2.00	0.00	0.00
2	3.42	–	3.10	4.03	0.00	0.00
3	1.34	3.03	–	3.21	1.60	4.56
4	1.00	1.28	2.12	–	2.56	1.00
5	0.00	1.00	1.07	1.80	–	1.00
6	0.00	0.00	1.00	1.00	1.00	–

Tabela 4.6: Tempos de Espera Médios Estimados.

Esta análise também permite inferir alguma informação interessante sobre o processo de velocidade de ventos na região de Settala: as transições entre estados de vento forte ocorrem num intervalo temporal curto, uma vez que o tempo médio de espera para a ocorrência de transição do estado 5 para o estado 6 ou do estado 6 para o estado 5 é de uma hora.

Em média, o tempo que o sistema demora a transitar para o estado 2, quando vem do estado 1, é de aproximadamente 8.56 horas. Curiosamente, é mais rápido a transitar para o estado 1 se provier do estado 2 - demora aproximadamente 3.43 horas. No entanto, relembrando os critérios climatológicos usados para a discretização em estados, é natural que os tempos de transição nos estados de vento fraco tenham uma banda temporal mais larga, uma vez que contêm velocidades de vento baixas e comuns na região, que podem variar de modo gradual ao longo do dia (note-se que variações de 1 m/s para 3 m/s ou de 4.5 m/s para 8.5 m/s, por exemplo, estão contempladas no mesmo estado - estados 1 ou 2, respetivamente - e por isso não são controladas). Em relação aos resultados temporais, realce-se que eram expectáveis, uma vez que a grande diferença entre a abordagem tradicional (de Markov) e a realizada na presente tese é a da consideração de informação temporal.

Antes de terminar a análise da modelação, deve salientar-se que, em relação às quantidades de interesse avaliadas nesta secção, não foram observadas melhorias significativas pela consideração do processo semi-markoviano de segunda ordem em relação ao de primeira ordem. No entanto, como já foi referido, o processo considerado é de segunda ordem nos estados e de primeira ordem no tempo. D'Amico *et al.* (2013) referem melhorias na

aproximação a dados reais por modelos semi-markovianos de segunda ordem nos estados e no tempo.

Globalmente, os resultados vistos permitem afirmar que **a modelação semi-markoviana aparenta fazer um melhor ajuste aos dados reais de velocidade máxima de vento horária do que a modelação markoviana**.

4.2 Previsão - Mecanismos de Exploração

Relativamente aos processos de **previsão de velocidade de vento**, face aos maus resultados obtidos nas tentativas de previsão horária, a partir deste momento a análise será feita sobre os dados diários máximos de velocidade de vento. A série a analisar é a da Figura 4.6.

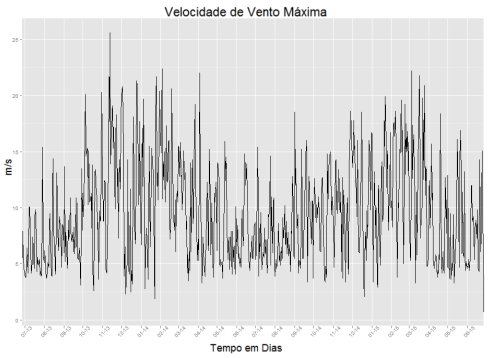


Figura 4.6: Série Temporal Diária.

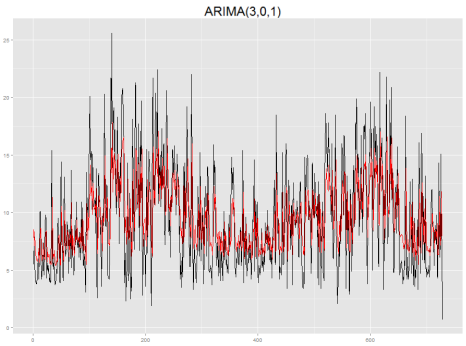


Figura 4.7: ARIMA(3,0,1).

Tentou modelar-se a série temporal por um ARIMA usando a instrução auto.arima do R (cujo algoritmo foi desenvolvido, apresentado e implementado em Hyndman e Khandakar (2007)). Esta função sugere o modelo ARIMA que melhor se adapta aos dados (com base em testes de raíz unitária para determinar o valor de d e com base nos critérios AIC, BIC ou AICc (AIC corrigido) para os outros dois parâmetros).

Segundo os testes de Dickey-Fuller e Phillips-Perron referidos na parte teórica, é rejeitada a hipótese de existir pelo menos uma raíz dentro do círculo unitário, para um nível de significância de 5%. O teste de KPSS não rejeita a hipótese de estacionariedade da série. Os valores-p para cada teste foram, respetivamente, 0.01, 0.01 e 0.1. No caso destes testes, não será possível haver concordância entre os resultados, uma vez que a hipótese de interesse é a hipótese nula num dos testes e a alternativa no outro. Como a não rejeição de uma hipótese nula não implica a sua aceitação, pode dizer-se apenas que, segundo os dois primeiros testes, a hipótese de não estacionariedade é rejeitada e segundo o terceiro teste, a estacionariedade não é rejeitada. Deste modo, supõe-se que o modelo ARIMA poderá ser aplicado a esta série sem necessidade de a diferenciar.

Face à previsão por séries temporais, note-se que a reta de referência no correlograma da ACF é dada por $y = -\frac{1}{n} \pm (\frac{2}{\sqrt{n}})$, e portanto depende do número de observações (n). Em seguida apresentam-se os correlogramas das autocorrelações totais e parciais, respetivamente, da série temporal em análise.

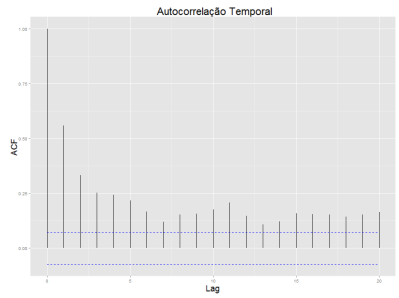


Figura 4.8: ACF.

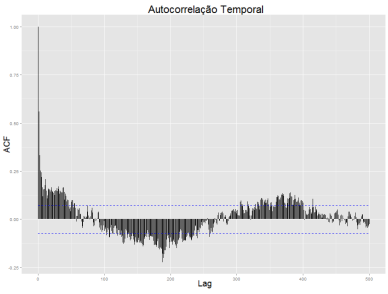


Figura 4.9: ACF (500 *lags*).

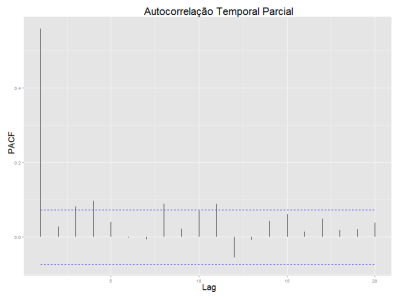


Figura 4.10: PACF.

A ACF não estabiliza abaixo da reta de referência nos primeiros *lags* observados. Para um maior número de *lags*, a ACF parece ter um comportamento análogo ao de uma onda criticamente amortecida. Como nem sempre a análise gráfica é correta e suficiente para justificar uma decisão, aliada ao facto de não se encontrarem referências com um suporte teórico sólido aos métodos gráficos de apoio à decisão do modelo, devem testar-se várias ordens e escolher a que apresentar melhores resultados. Neste caso, como os dois modelos referidos nesta secção foram utilizados sobretudo como **mecanismos de exploração** da série temporal, esse estudo não foi feito. O modelo escolhido pela instrução usada foi um $\mathbf{ARIMA(3,0,1)}$ com média diferente de zero e AIC= 3966.32, AICc= 3966.43 e BIC= 3993.86. A análise dos resíduos mostra que a hipótese de independência dos resíduos não é rejeitada (até 10 *lags*), para o teste de Ljung-Box [Ljung e Box (1978)].

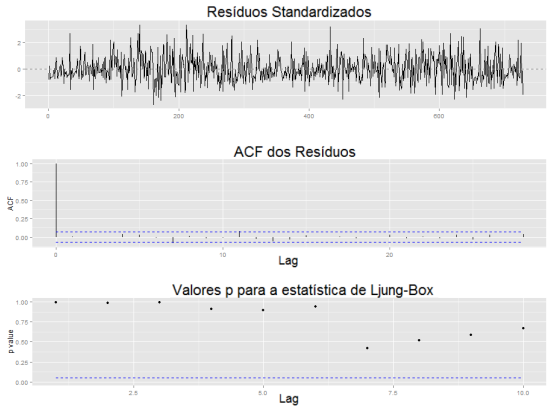


Figura 4.11: Análise de Resíduos do Modelo ARIMA.

Para fornecer uma maior consistência aos resultados, utilizou-se o mesmo teste de Ljung-Box para testar a independência serial na série original, que por sua vez foi rejeitada (note-se que este teste faz uma aproximação à série por modelos ARMA). Foi também utilizado um teste não paramétrico de **Kruskall-Wallis** para testar a igualdade de distribuições empíricas entre a variável Dia e variável Velocidade de Vento, que não foi rejeitada, tendo obtido um valor-p de 0.48. Apesar de os resultados aparentarem aproximar um sinal de ruído branco (tal como se espera), não se deve esquecer que os modelos ARIMA usados são lineares e, existindo padrões não-lineares, podem ser captados pelo modelo e adulterar as previsões por ele feitas.

Por esse motivo, foi explorada uma outra técnica, designada por Análise Espetral Singular (SSA) [Elsner e Tsonis (2013)], que não obriga ao conhecimento sobre o modelo paramétrico da série temporal. É composta pelas etapas de **decomposição** - por sua vez composta nas etapas de **embutimento** e **decomposição em valores singulares** - e **reconstrução** - por sua vez composta pelo **agrupamento de triplos próprios** e **média da diagonal**. Assumindo que a série estudada é decomposta na soma de tendência, componentes oscilatórias e ruído, este método só obriga à definição de dois parâmetros: tamanho da janela L (cujo comprimento ótimo é dado por $L_{\max} = \frac{N}{2}$, onde N é o número de observações da série; para se conseguir alcançar separabilidade suficiente das componente, aconselha-se a escolha de um valor de L proporcional ao período de sazonalidade dos dados [Golyandina *et al.* (2001)]) a usar e triplos próprios a agrupar. Estes conceitos serão explicados de seguida.

Procedimento SSA:

Cada triplo é um vetor próprio, um vetor fator e um vetor singular. Na etapa sequencial de SSA, inicialmente faz-se a extração de tendência usando o primeiro vetor próprio, tal como descrito por Golyandina e Korobeynikov (2014). Em seguida extraem-se as componentes aleatórias dos resíduos. Para tal, devem agrupar-se os triplos próprios com valores singulares próximos, uma vez que a quebra no espetro dos valores próprios permite detetar uma sequência lentamente decrescente de valores singulares produzida por um sinal de ruído branco. Antes de decidir quais os triplos próprios a agrupar, deve analisar-se a matriz de correlações ponderadas entre as componentes obtidas na separação de valores próprios (componentes reconstruídas).

Deve analisar-se ainda o gráfico dos vetores próprios sucessivos, agrupando os triplos próprios associados a polígonos regulares. Note-se que este mecanismo funciona com base no conceito de *separabilidade*, defendendo que as diferentes componentes da série são identificáveis e separáveis, permitindo decompor a mesma. Habitualmente toma-se para a extração de tendência um valor de L proporcional ao período da série, mas como não se identifica (visualmente) nenhum período na série original, considerar-se-á, nas duas fases do procedimento, que o tamanho da janela será $L = \frac{728}{2} = 364$ (dias).

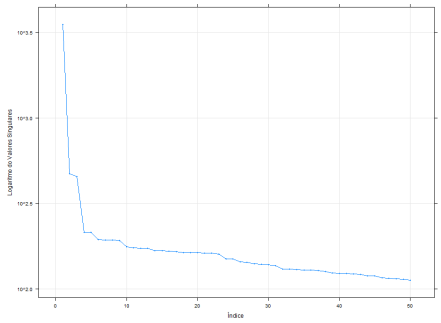


Figura 4.12: SSA - Valores Singulares.

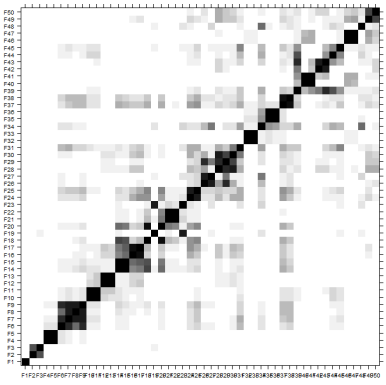


Figura 4.13: SSA - Matriz de Correlações entre as Componentes.

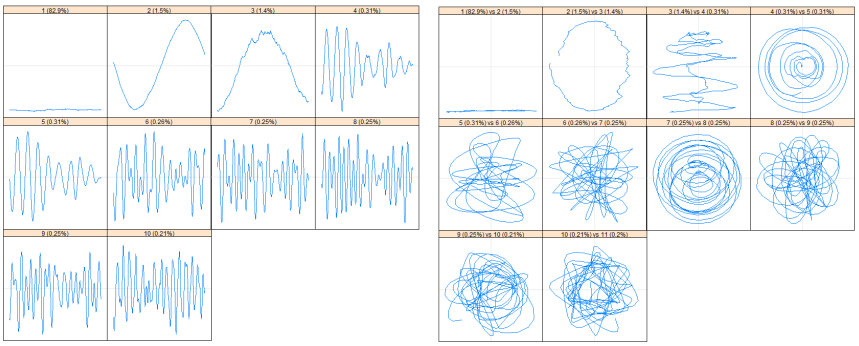


Figura 4.14: SSA - Vetores Próprios e Pares de Vetores Próprios.

Embora a análise gráfica não tenha sido trivial, uma vez que apenas o par (4, 5) é claramente identificável (porque estes dois vetores estão emparelhados na Figura 4.12, apresentam igual comportamento nos vetores próprios da Figura 4.14 e correspondem às poucas componentes claramente identificadas por uma alta correlação na Figura 4.13), pela análise gráfica destas três componentes de decisão, os triplos escolhidos foram os seguintes: (4, 5) e (7, 8).

Observe-se também que a SSA não identifica periodicidade na série, uma vez que não existem polígonos regulares definidos no gráfico de Pares de Vetores Próprios da Figura 4.14. Esta informação é útil para concluir algo que já se desconfiava: a série aparenta ser definida por diversas componentes oscilatórias, cada uma com a sua quota parte de participação na série original, mas não parece existir periodicidade em nenhuma das componentes dos triplos selecionados, pelo que a separabilidade da série pode ser feita (usando os triplos próprios e a tendência entretanto extraídos), mas não se consegue fazer uma interpretação física das mesmas.

Uma vez feito o agrupamento, prosseguiu-se com a reconstrução da série, que se encontra na Figura 4.15, onde se vê a série original, a remoção da tendência (a azul) na primeira etapa da SSA sequencial e que centra a série, e a reconstrução da mesma com base nos triplos próprios definidos anteriormente.

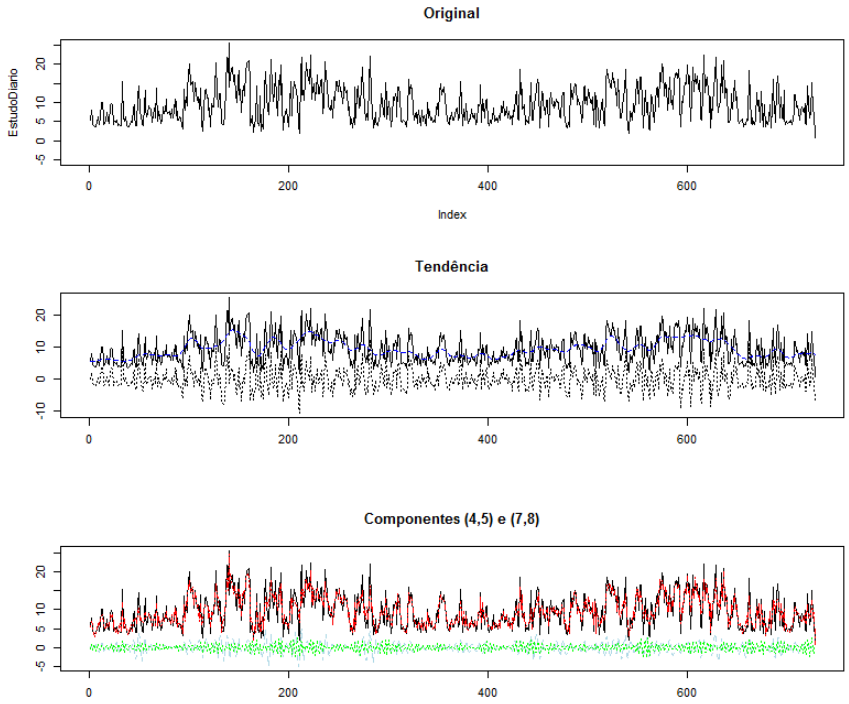


Figura 4.15: SSA - Reconstrução da Série.

Com esta informação, é possível prever dados com base no algoritmo (de recorrência) descrito no capítulo 5 da obra de Golyandina *et al.* (2001).

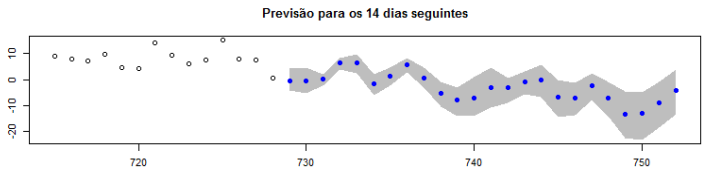


Figura 4.16: SSA - Previsão.

4.3 Previsão - Aplicação e Comparação de Resultados

Para a previsão, tal como referido no Capítulo 3, foram usadas Redes Neuronais Artificiais, Máquinas de Suporte Vetorial e Árvores de Regressão (cujas instruções em R são dadas por nnet, svm e rpart, respetivamente). Relembrando, a validação cruzada é usada para avaliar a capacidade de predição/previsão do modelo com uns determinados parâmetros (o k-fold é o método de partição dos dados e é independente do método que se usa para a separação inicial; é usado para que haja reamostragem e o treino não seja feito sempre sobre os mesmos dados). Usaram-se 10 folds para a validação, o que significa que o algoritmo na validação cruzada é treinado com 90% do treino e testado nos restantes 10% do treino. O erro de previsão final do modelo, nesta etapa, é dado pelo MAPE no conjunto de treino (embora também tenha sido calculado o erro MAD). Com a informação que se obtém do passo anterior, determinam-se os parâmetros que obtêm um menor erro no treino (isto para cada um dos métodos estudados). Nesta etapa, já se dispõe dos valores de desempenho de cada modelo. Comparando-os, existe um **modelo vencedor** e é esse modelo que vai ser usado para fazer previsões no conjunto de teste.

Os conjuntos de treino e teste iniciais foram definidos de duas formas:

- **Aleatória**: separação aleatória da amostra em dois conjuntos: treino - 70% e teste - 30%.
- **Sequencial**: separação da amostra em dois conjuntos: treino - primeiros 70% das observações da base de dados - e teste - restantes 30% das observações da base de dados. Uma vez que esta separação é organizada no tempo, é possível comparar os modelos acima vistos com o modelo ARIMA, usando a técnica de *Sliding Window*.

Far-se-á de seguida a análise dos resultados obtidos.

Análise dos resultados da separação inicial aleatória:

Os resultados da validação cruzada para cada modelo foram os seguintes:

- Árvores de regressão:

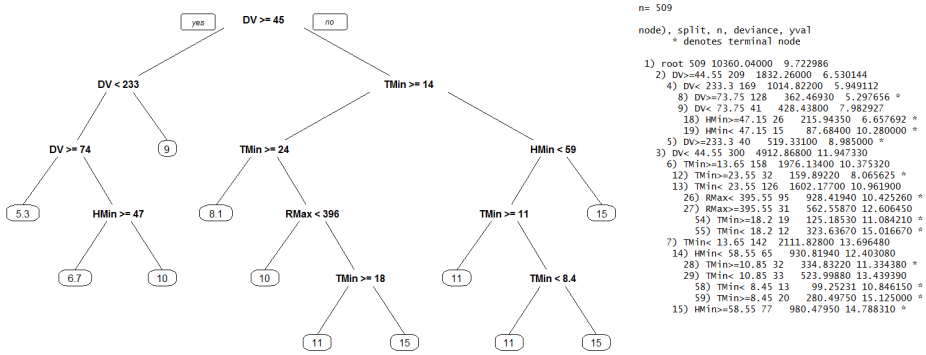


Figura 4.17: Previsões Numéricas por Árvores de Regressão.

O erro percentual médio absoluto deste método, estimado por validação cruzada, foi de 32.24 % e o erro absoluto médio foi de 2.72. As previsões numéricas mostraram que a Direção de Vento é uma variável fundamental no processo de decisão, o que justifica o facto de, em grande parte dos estudos eólicos, estas duas variáveis serem estudadas em conjunto. Para direções de vento iguais ou superiores a 45^o, prevêem-se, no máximo, ventos de 10 m/s, ao passo que, para direções inferiores a 45^o, se podem prever ventos até 15 m/s. Para se alcançar este último valor, por exemplo, seria necessário que, na região de Settala, se verificasse uma das seguintes condições (climáticas):

1. Direção de Vento inferior a 45^o medidos relativamente à direção de referência, Temperatura Mínima superior a 14^oC e inferior a 18^oC, bem como Radiação Global Máxima superior a 396 W/m^2;

2. Direção de Vento inferior a 45^o medidos relativamente à direção de referência, Temperatura Mínima inferior a 14^oC e Humidade Relativa Mínima superior a 59%;

3. Direção de Vento inferior a 45^o medidos relativamente à direção de referência, Temperatura Mínima inferior a 11^oC e superior a 8.4^oC.

Tomando mais atenção à variável Direção de Vento, note-se que até agora apenas se sabe que é expressa em graus e a sua interpretação é feita de modo relativo (como visto acima). No entanto, após alguma pesquisa, encontrou-se, no site da Organização Mundial de Meteorologia, a seguinte informação: "Os ventos são denominados a partir da direção de onde sopram" e as suas direções de referência são as seguintes:

N	NNE	NE	ENE	E	ESE	SE	SSE	S	SSW	SW	WSW	W	WNW	NW	NNW
0^o	22.5^o	45^o	67.5^o	90^o	112.5^o	135^o	157.5^o	180^o	202.5^o	225^o	247.5^o	270^o	293.5^o	315^o	337.5^o

Tabela 4.7: Valores de Referência na Medição da Direção do Vento.

Conclui-se, com esta informação, que segundo a previsão executada pelas árvores de regressão, os ventos mais fortes (previstos) são determinados por Temperaturas Mínimas inferiores a 18^oC e ventos que sopram de Norte (N) a Nordeste (NE).

- Redes neuronais artificiais: Por cada combinação de parâmetros (*size*, *decay*) da instrução nnet em R (onde *size* é o número de neurónios na camada escondida e *decay* é um parâmetro de decaimento dos pesos atribuídos na rede), são calculados o MAPE médio, MAD médio e o desvio padrão para MAPE e MAD. Tal como já foi referido, o critério de decisão será o valor médio do erro MAPE.

size	***decay***	$\hat{\mu}_{\mathbf{MAPE}}$	$\hat{\sigma}_{\mathbf{MAPE}}$	$\hat{\mu}_{\mathbf{MAD}}$	$\hat{\sigma}_{\mathbf{MAD}}$
1	0.1	48.13%	9.12%	3.62	0.67
2	0.1	41.89%	10.29%	3.24	0.62
3	0.1	50.58%	6.99%	3.70	0.36
4	0.1	48.61%	10.25%	3.59	0.42
5	0.1	41.67%	12.67%	3.11	0.47
1	0.01	50.92%	6.84%	3.75	0.37
2	0.01	50.92%	6.83%	3.75	0.37
3	0.01	50.12%	8.43%	3.69	0.45
4	0.01	49.71%	8.95%	3.69	0.47
5	0.01	50.92%	6.84%	3.75	0.37
1	0.05	41.43%	12.58%	3.21	0.78
2	0.05	45.82%	10.47%	3.46	0.74
3	0.05	45.68%	10.86%	3.46	0.64
4	0.05	46.62%	11.92%	3.51	0.72
5	0.05	48.09%	10.12%	3.58	0.49

Tabela 4.8: Informação de Decisão Proveniente da Validação Cruzada nas ANN.

A informação dada na Tabela 4.8 mostra que o menor erro percentual médio absoluto, estimado por validação cruzada, foi registado para os parâmetros *size*= 1 e *decay*= 0.05.

- Regressão com máquinas de suporte vetorial: Por cada combinação de parâmetros (*cost*, *gamma* e *epsilon*) da instrução svm em R (onde *cost* é uma constante de regularização do processo, *gamma* é um parâmetro da função núcleo e *epsilon* é designada por *insensitive-loss function*), são calculados os mesmos erros já introduzidos. Tal como já foi referido na contextualização teórica da tese, o critério de decisão será o valor médio do erro MAPE. Uma vez que foram testadas várias combinações de valores para os parâmetros desta função ($cost \in \{0.5, 1.5, ..., 6\}$, $gamma \in \{0.1, 0.3, ..., 1\}$, $epsilon \in \{0.1, 0.2, ..., 1\}$), apresenta-se o resultado final (evitando a apresentação de uma tabela exaustiva com a apresentação da informação fundamental: a dos valores dos parâmetros que obtiveram menor erro médio MAPE na validação cruzada).

cost	***gamma***	**epsilon**	$\hat{\mu}_{\mathbf{MAPE}}$	$\hat{\sigma}_{\mathbf{MAPE}}$	$\hat{\mu}_{\mathbf{MAD}}$	$\hat{\sigma}_{\mathbf{MAD}}$
2.5	0.3	0.1	27.51%	5.42%	2.33	0.37

Tabela 4.9: Informação de Decisão Proveniente da Validação Cruzada nas SVM.

Comentário: Do processo de validação cruzada, já se têm os menores valores de erro médio para cada modelo. Consequentemente, o modelo que vai ser usado no conjunto de teste é o modelo de **regressão com máquinas de suporte vetorial**, uma vez que apresentou o menor valor $\hat{\mu}_{\mathrm{MAPE}}$.

O treino é usado para escolher o melhor modelo. Os valores que se vão prever são os do conjunto de teste. Pode fazer-se uma última avaliação do mesmo, desenhando os valores previstos e reais para o conjunto de teste e pode complementar-se essa análise com o valor dos erros. Antes de o fazer, mostram-se os resultados obtidos pela separação sequencial dos dados amostrais.

Análise dos resultados da separação inicial sequencial:

- Árvores de Regressão:

$\hat{\mu}_{\mathbf{MAPE}}$	$\hat{\mu}_{\mathbf{MAD}}$
30.02%	2.46

Tabela 4.10: Erros nas Árvores de Regressão, no Caso Sequencial.

- Redes neuronais artificiais:

size	***decay***	$\hat{\mu}_{\mathbf{MAPE}}$	$\hat{\sigma}_{\mathbf{MAPE}}$	$\hat{\mu}_{\mathbf{MAD}}$	$\hat{\sigma}_{\mathbf{MAD}}$
1	0.1	40.97%	10.96%	3.03	0.6
2	0.1	45.84%	8.80%	3.38	0.49
3	0.1	48.08%	7.85%	3.53	0.44
4	0.1	45.39%	9.37%	3.38	0.56
5	0.1	43.03%	10.98%	3.20	0.59
1	0.01	49.52%	4.76%	3.60	0.23
2	0.01	49.52%	4.76%	3.60	0.24
3	0.01	49.52%	4.76%	3.60	0.24
4	0.01	49.51%	4.77%	3.60	0.24
5	0.01	45.71%	10.44%	3.33	0.52
1	0.05	39.47%	10.79%	3	0.66
2	0.05	47.78%	6.47%	3.48	0.38
3	0.05	42.43%	11.49%	3.10	0.68
4	0.05	45.34%	9.77%	3.35	0.57
5	0.05	48.12%	5.77%	3.49	0.34

Tabela 4.11: Erros nas Redes Neuronais Artificiais, no Caso Sequencial.

- Regressão com máquinas de suporte vetorial:

cost	***gamma***	**epsilon**	$\hat{\mu}_{\mathbf{MAPE}}$	$\hat{\sigma}_{\mathbf{MAPE}}$	$\hat{\mu}_{\mathbf{MAD}}$	$\hat{\sigma}_{\mathbf{MAD}}$
1.5	0.3	0.1	26.68%	4.19%	2.23	0.24

Tabela 4.12: Erros nas SVM, no Caso Sequencial.

- ARIMA: gráfico com os erros MAPE em cada combinação (p,0,q) testada:

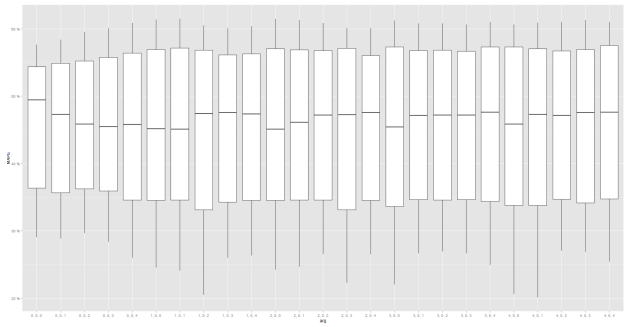


Figura 4.18: Erros nos modelos ARIMA, no Caso Sequencial.

Comentário: Para se determinar a melhor ordem no caso ARIMA, analisa-se o valor do erro MAPE em cada um dos *steps* (isto é, número de dias analisado em cada janela deslizante no método *Sliding Window*), para cada uma das combinações testadas. Posteriormente seleciona-se a menor média de MAPE desses valores e procura-se a combinação ARIMA(p,0,q) que lhe está associada. Neste caso, a ordem com menor erro foi a dada por um ARIMA(1,0,1), com erro MAPE médio de 44.9%.

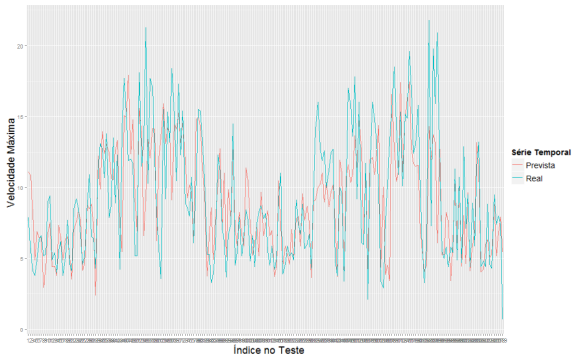


Figura 4.19: Previsão por SVM aleatória.

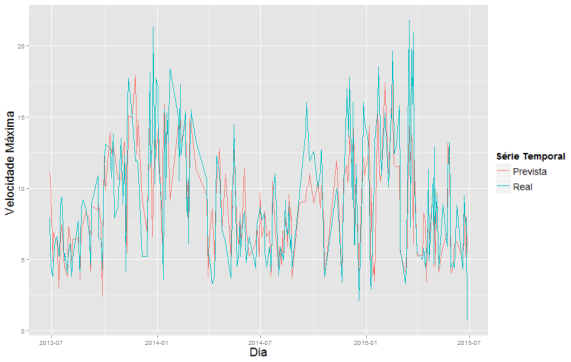


Figura 4.20: Previsão por SVM sequencial.

Concluindo, o modelo SVM foi o escolhido, nas duas situações de partição inicial da amostra, para fazer previsão no conjunto de teste. Não parece haver grandes diferenças, numéricas ou gráficas, entre os resultados dos dois modelos. Tendo em conta a estrutura temporal dos dados e a ordem de grandeza dos erros, parece natural a escolha do modelo sequencial.

-	**MAPE**	**MAD**
Caso Aleatório	29.94%	2.22915
Caso Sequencial	30.94%	2.23455

Tabela 4.13: Erros no Conjunto de Teste.

Capítulo 5

Modelo Generalizado de Sparre Andersen

Neste capítulo serão feitas referências teóricas à atividade seguradora, em particular ao modelo de Sparre Andersen (GMSA), que pode ser usado na sua modelação. É importante referir que uma modelação adequada dos eventos que poderão dar origem a indemnizações extremas é essencial para a atividade de uma Companhia de Seguros, porque lhe permite modelar corretamente essas indemnizações, definindo um prémio apropriado, modelando adequadamente o resseguro e determinando o capital necessário para minimizar a probabilidade de insolvência da Companhia, em particular na área de Seguros de Não Vida cujas indemnizações, em caso de sinistro, são muitos altas.

Deve salientar-se que o modelo que se apresenta em seguida é uma referência teórica e explicada de um modo que se espera mais intuitivo do que o original, mas que já foi definida por Drekic e Mera (2011) e Alfa e Drekic (2007), baseada em processos markovianos. Esta abordagem refere-se aos conceitos matemáticos (tentando uma explicação acessível) que permitem a configuração do problema, não se focando tanto na algoritmia como os autores referidos o fizeram.

A vantagem da generalização que este modelo apresenta, permite que, ao longo das próximas páginas se possa presumir que estamos perante um processo de modelação de quantidades de ruína de uma seguradora que se dedica apenas a danos por velocidades de vento, ou então que se dedica a um outro qualquer ramo segurador, claramente orientado, uma vez que o sistema em estudo deixa de ser o dano (ou a sua causa), mas a entidade que gere os acontecimentos (graves) por ele desencadeadas.

5.1 Quantidades de Ruína e Limiar de Prémio

O GMSA é utilizado sobretudo em Teoria de Risco, em alternativa ao modelo clássico. O principal objetivo do uso deste modelo é a modelação das principais quantidades associadas à ruína: *instante de ruína*, *défice na ruína* e *superávit imediatamente anterior à ruína*. Esta última corresponde à quantia exata de que a seguradora dispunha imediatamente antes da ocorrência de ruína.

Assume-se que o processo do número de indemnizações, dado por $\{N_t : t = 0, 1, ...\}$, é um processo de renovamento modificado em tempo discreto, com tempos entre indemnizações positivos e independentes, onde W_1 é a duração do tempo 0 até ao instante da primeira indemnização e W_i o tempo entre as (i-ésima-1) e i-ésima indemnização. Há que notar que (W_1,W_2,...) forma uma sequência de variáveis aleatórias independentes e identicamente distribuídas com função de probabilidade dada por

$$a_j = P(W_i = j) \quad j = 1, 2, ..., n_a \quad (n_a < \infty)$$

e correspondente função de sobrevivência

$$A_j = P(W_i > j) = 1 - \sum_{k=1}^{j} a_k.$$

Assume-se também que a distribuição do tempo entre indemnizações tem suporte finito.

No contexto da análise de risco feita numa seguradora, segundo o modelo clássico, os tempos entre indemnizações formam uma sequência de variáveis aleatórias independentes com distribuição exponencial de parâmetro λ. Assim, o número de indemnizações segue um processo de Poisson, o que, devido à perda de memória que caracteriza a a exponencial, faz com que as variáveis "tempo até à primeira indemnização" (W_1) e "instante da primeira indemnização" (T_1) tenham a mesma distribuição. No caso do GMSA, a forma como este processo é conduzido é geral, na medida em que a distribuição dos tempos entre indemnizações é arbitrária (e portanto a distribuição de W_1 não será necessariamente a mesma de T_1, a não ser que tenha ocorrido alguma indemnização no instante 0).

Ora, como W_1 será tratado como um caso à parte, uma vez que não existe informação sobre o que se passou antes da ocorrência da primeira indemnização, há duas considerações feitas habitualmente:

- assume-se que ocorreu uma indemnização antes de 0, pelo que W_1,W_2,... passam a ter a mesma distribuição
- assume-se o modelo mais geral, em que W_1 segue uma qualquer função de probabilidade $r_j = P(W_1 = j) \quad j = 1, 2, ..., n_r \quad (n_r < \infty)$ e correspondente função de sobrevivência $R_j = P(W_1 > j) = 1 - \sum_{k=1}^{j} r_k$

Assim, o principal processo de risco analisado é o seguinte:

$$U_t = u + \sum_{i=0}^{t-1} p_i - \sum_{i=1}^{N_t} Y_i$$

onde U_t representa o montante do superávit da seguradora no instante t, ou seja, a reserva de risco de uma carteira no instante t. Este processo é, na verdade, um balanço de contas entre o que entra de prémios na seguradora e o que sai para pagamento de indemnizações,

não esquecendo a reserva inicial dada por $u = U_0 \in \mathbb{N}$.

Da forma como se define, U_t representa o montante de superávit no final do intervalo $(t-1, t]$. Em relação a este intervalo em particular, assume-se que os prémios são recebidos em $(t-1)^+$ e as indemnizações pagas em t^-, garantindo um balanço real em t.

É agora adicionada ao *superávit* da seguradora uma quantia $\Upsilon \in \mathbb{Z}^+$ que afeta o prémio recebido num certo instante e que funciona como um limiar entre um prémio constante e um prémio aleatório (isto porque, a partir de um determinado montante de reserva, a companhia já tem segurança financeira suficiente para flexibilizar os prémios dos seus clientes). Assim, define-se que um prémio aleatório

$$p_t = \begin{cases} c & \text{se } U_t < \Upsilon \\ X_t & \text{se } U_t \geq \Upsilon \end{cases}$$

com $c \in \mathbb{Z}^+$ como o prémio sem encargos e X_t como um prémio aleatório recebido em t, com

$$d_i = P(X_t = i), \quad i = c_1, c_1 + 1, ..., c_2, \quad \sum_{i=c_1}^{c_2} d_i = 1.$$

c_1, c_2 são, respetivamente, os valores mínimo e máximo do suporte da distribuição de X_t, com $c_1, c_2 \in \{0, 1, ..., c\}$ e $c_1 \leq c_2$. Assim, a quantia $c - p_t$ pode ser interpretada como a participação de resultados, muitas vezes referida no mundo segurador.

Sejam $\{Y_1, Y_2, Y_3...\}$ os montantes individuais das indemnizações que a seguradora tem de pagar - variáveis aleatórias positivas *i.i.d.* com função densidade de probabilidade

$$\alpha_j = P(Y_i = j) \quad j = 1, 2, ..., m_\alpha$$

e correspondente função de sobrevivência

$$\Lambda_j = 1 - \sum_{k=1}^{j} \alpha_k \quad (m_\alpha \leq \infty).$$

Consideram-se as seguintes quantidades de ruína:

- o **instante de ruína**, escrito como

$$T = min\{t \in \mathbb{Z}^+ \mid U_t < 0\}$$

  e $T = \infty$ se $U_t \geq 0 \quad \forall t \in \mathbb{Z}^+$

- se a ruína ocorrer, o **défice na ruína** é dado por $\mid U_T \mid$

- se a ruína ocorrer, o **superávit imediatamente anterior à ruína** é dado por

$$U_{T^-} = U_{T-1} + p_{T-1}$$

Portanto $T = \infty$ se $m_\alpha \leq min\{c, \Upsilon + c_1\}$. Por outro lado, se $m_\alpha > min\{c, \Upsilon + c_1\}$, então

$$\mid U_T \mid \in \{1, 2, ..., m_\alpha - min\{c, \Upsilon + c_1\}\} \quad e$$

$$U_{T^-} \in \{min\{c, \Upsilon + c_1\}, min\{c, \Upsilon + c_1\} + 1, ..., m_\alpha - 1\}\}.$$

O que a notação acima definida mostra é que, se o valor das indemnizações a pagar aos segurados é superior à entrada de dinheiro que existe na seguradora, proveniente do pagamento de prémios, a ruína é certa e o défice da ruína pode ir desde uma unidade monetária até à diferença entre o valor que a companhia teve de pagar em indemnizações e o montante de que dispunha até esse acontecimento. Do mesmo modo, o superávit imediatamente anterior à ruína pode ir desde o valor máximo de que a seguradora dispunha antes da indemização (caso $m_\alpha = min\{c, \Upsilon + c_1\}$) até a uma unidade monetária a menos que a necessária para pagar essa mesma indemnização.

Uma vez definidas as quantidades de ruína a modelar, devem introduzir-se as probabilidades conjuntas que lhes estão associadas e onde se pretende chegar para que a modelação seja possível. Sejam:

- $\omega_{n,i}(u) = P(T = n, U_{T^-} = i, \mid U_0 = u)$
- $\phi_{n,j}(u) = P(T = n, \mid U_T \mid = j \mid U_0 = u)$
- $\psi_{n,i,j}(u) = P(T = n, U_{T^-} = i, \mid U_T \mid = j \mid U_0 = u)$

Mais à frente, será visto o modo de estimar as probabilidades associadas com o uso das variáveis aleatórias $\mid U_T \mid$ e U_{T^-}.

5.2 Formulação do Modelo

Daqui em diante, seja $W = W_i$ arbitrário com $i = 2, 3, ...$ e

$$\tau_j = P(W > j \mid W > j - 1) = \frac{A_j}{A_{j-1}} \qquad (\tau_1 = A_1 \text{ e } \tau_{n_a - 1} = 0)$$

a probabilidade do tempo entre indemnizações ser superior a j unidades de tempo, sabendo que até ao instante de análise anterior, a indemnização não tinha ocorrido. Considera-se ainda

$$S = \begin{pmatrix} 0 & \tau_1 & 0 & \cdots & 0 \\ 0 & 0 & \tau_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & \ddots & 0 & \tau_{n_a-1} \\ 0 & 0 & \ddots & 0 & 0 \end{pmatrix}, \boldsymbol{s} = \begin{pmatrix} 1 - \tau_1 \\ 1 - \tau_2 \\ \vdots \\ 1 - \tau_{n_a-1} \\ 1 \end{pmatrix} \text{ e } \boldsymbol{e}_1 = \big(1, 0, ..., 0\big)$$

tais que $a_j = P(W_i = j) = \boldsymbol{e}_1 S^{j-1} \boldsymbol{s}$. Esta definição foi dada e demonstrada por Wu e Li (2008).

O objetivo intermédio deste modelo é a construção do processo bivariado

$$\{(U_t, L_t) : t = k, k+1, ...\}.$$

em que U_t representa - tal como visto anteriormente - o *superávit* da seguradora no instante t e L_t denota um contador de tempo em t que mede o "tempo que falta" até à próxima indemnização. Note-se ainda que a componente U corresponde ao nível do processo, ao passo que a componente L corresponde à fase do processo, que segue uma relação Markoviana:

$$(U_{t+1}, L_{t+1}) = \begin{cases} (U_t + p_t, L_t + 1) & \text{se não existe indemnização em } (t+1)^+ \\ (U_t + p_t - Y, 1) & \text{se existe uma indemnização de } Y \text{ em } (t+1)^+ \end{cases}$$

É agora possível analisar a matriz que contém as probabilidades de transição associadas a esta **cadeia de Markov** (com um espaço de estados dado por $\Delta = \mathbb{Z}$ x $\{1, 2, ..., n_a\}$):

$$\begin{array}{c} \\ \vdots \\ -1 \\ 0 \\ 1 \\ \vdots \\ \Upsilon - 2 \\ \Upsilon - 1 \\ \Upsilon \\ \Upsilon + 1 \\ \vdots \end{array} \begin{array}{c} \begin{array}{cccccccc} \cdots & -1 & 0 & 1 & \cdots & \Upsilon - 2 & \Upsilon - 1 & \Upsilon & \cdots \end{array} \\ \begin{pmatrix} \ddots & \vdots & \vdots & \vdots & \cdots & \vdots & \vdots & \vdots & \vdots \\ \cdots & B_c & B_{c-1} & B_{c-2} & \cdots & B_{c-\Upsilon+1} & B_{c-\Upsilon} & B_{c-\Upsilon-1} & \cdots \\ \cdots & B_{c+1} & B_c & B_{c-1} & \cdots & B_{c-\Upsilon+2} & B_{c-\Upsilon+1} & B_{c-\Upsilon} & \cdots \\ \cdots & B_{c+2} & B_{c+1} & B_c & \cdots & B_{c-\Upsilon+3} & B_{c-\Upsilon+2} & B_{c-\Upsilon+1} & \cdots \\ \cdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots \\ \cdots & B_{\Upsilon+c-1} & B_{\Upsilon+c-2} & B_{\Upsilon+c-3} & \cdots & B_c & B_{c-1} & B_{c-2} & \cdots \\ \cdots & A_{\Upsilon+1} & A_\Upsilon & A_{\Upsilon-1} & \cdots & A_2 & A_1 & A_0 & \cdots \\ \cdots & A_{\Upsilon+2} & A_{\Upsilon+1} & A_\Upsilon & \cdots & A_3 & A_2 & A_1 & \cdots \\ \cdots & A_{\Upsilon+3} & A_{\Upsilon+2} & A_{\Upsilon+1} & \cdots & A_4 & A_3 & A_2 & \cdots \\ \cdots & \vdots & \vdots & \vdots & \cdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix} \end{array}$$

onde cada elemento da matriz é uma de duas matrizes por blocos, definidas como

$$B_i = \begin{cases} O_{n_a} & \text{if } i \in \mathbb{Z}^- \\ S & \text{se } i = 0 \\ (\boldsymbol{s}\boldsymbol{e}_1)\alpha_i & \text{se } i \in \mathbb{Z}^+ \end{cases} \quad \text{e} \quad A_i = \sum_{j=c_1}^{c_2} d_j B_{i+j}$$

Note-se que este processo é uma cadeia de Markov dupla, cuja matriz P apresenta blocos finitos de dimensão n_a x n_a. Como cada elemento da matriz corresponde a outra matriz, a Figura 5.1 seguinte tenta ilustrar a estrutura da matriz, para que haja uma melhor interpretação da mesma.

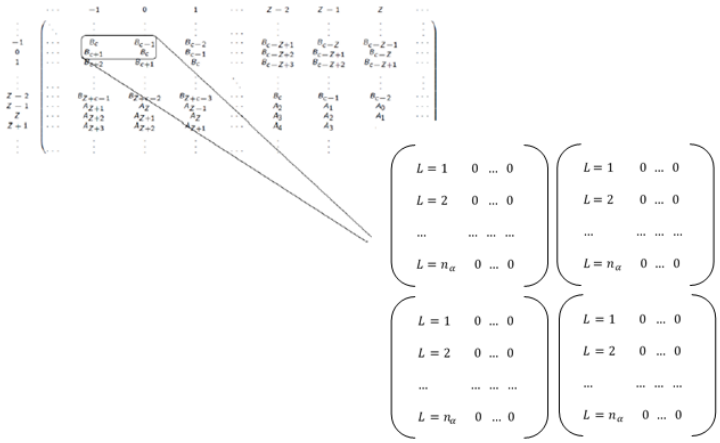


Figura 5.1: Estrutura Matricial de Transições no GMSA.

Considere-se agora um espaço de estados onde $\Delta_1 = \mathbb{N}$ x $\{1, 2, ..., n_a\}$ e $\Delta_2 = \mathbb{Z}^-$ x $\{1, 2, ..., n_a\}$ são tais que

$$\Delta_1 \cap \Delta_2 = \emptyset.$$

Tomem-se agora duas matrizes de transição (sub-matrizes de P)

$$C : \Delta_1 \to \Delta_1 \text{ e } D : \Delta_1 \to \Delta_2$$

que mapeiam, respetivamente, estados de *não ruína* em Δ_1 e estados de *ruína* em Δ_2 (C é o quadrante inferior direito de P e D é o quadrante inferior direito de P horizontalmente invertida).

Para o cálculo das quantidades relacionadas com a ruína e vistas anteriormente, terão de ser acrescentadas algumas definições, necessárias mais à frente. Sejam:

- $z_t = \begin{cases} min\{i \in \{1, 2, ..., t\} \mid u + c(i-1)\} & \text{se } u < T \\ 0 & \text{se } u \geq T \end{cases}$

- $d_{m,n} = \begin{cases} \sum_{j=c1}^{c2} d_{j,1} d_{m-j,n-1} & \text{se } m = nc_1, nc_1 + 1, ..., nc_2 \\ 0 & \text{caso contrário} \end{cases}$

lembrando que

$$d_{m,0} = \delta_{m,0} \text{ (onde } \delta \text{ é a função de Kronecker)}$$

$$d_{m,1} = \begin{cases} d_m & \text{se } m = nc_1, nc_1 + 1, ..., nc_2 \\ 0 & \text{caso contrário} \end{cases}$$

Como $W_1 = k$, $\boldsymbol{b}^{(k)}$ é o vector-linha com as probabilidades iniciais de estar nos estados de Δ_1 e é dada por

$$\sum_{m=(k-z_k)c1}^{(k-z_k)c2} d_{m,k-z_k}(\alpha_{u+cz_k+m}\boldsymbol{e}_1, \alpha_{u+cz_k+m-1}\boldsymbol{e}_1, ..., \alpha_2\boldsymbol{e}_1, \alpha_1\boldsymbol{e}_1, \boldsymbol{0}, \boldsymbol{0}, ...).$$

Agora são definidos dois vectores-coluna fundamentais:

- $\boldsymbol{g}_n^{(k)} = (\boldsymbol{g}_{n,0}^{(k)}, \boldsymbol{g}_{n,1}^{(k)}, \boldsymbol{g}_{n,2}^{(k)}, ...) = \boldsymbol{b}^{(k)} C^n \quad n \in \mathbb{N}$, correspondente à probabilidade de transição de um estado de não ruína para outro de não ruína em k unidades de tempo;

- $\boldsymbol{h}_n^{(k)} = (\boldsymbol{h}_{n,-1}^{(k)}, \boldsymbol{h}_{n,-2}^{(k)}, \boldsymbol{h}_{n,-3}^{(k)}, ...) = \boldsymbol{b}^{(k)} C^{n-1} D \quad n \in \mathbb{Z}^+$, que corresponde à probabilidade de transição de um estado de não ruína para um estado de ruína em k unidades de tempo.

Lembrando que $\boldsymbol{h}_{n,-j}^{(k)} = (\phi_{n,j}^{(k)}(u), 0, ..., 0)$ com $n \in \mathbb{Z}^+$

$$\phi_{n,j}^{(k)}(u) = P(T = k+n, \mid U_T \mid = j \mid U_k \in \Omega_k)$$

$j = 1, 2, ..., m_\alpha - min\{c, \Upsilon + c_1\}$, $\Omega_k = \{0, 1, ..., u + cz_k + c_2(k - z_k)\}$

obtém-se

$$\phi_{n,j}^{(k)}(u) = \boldsymbol{h}_{n,-j}^{(k)} \boldsymbol{e}_1^{\mathrm{T}}.$$

Aplicando o mesmo raciocínio, obtém-se uma representação para

$$\psi_{n,i,j}^{(k)}(u) = P(T = k+n, U_T = i, \mid U_T \mid = j \mid U_k \in \Omega_k)$$

de onde segue imediatamente que

$$\psi_{n,i,j}^{(k)}(u) = (\boldsymbol{g}_{n-1,i-p_{k+n-1}}^{(k)} \boldsymbol{s}) \alpha_{i+j}.$$

Uma vez que é utilizado um limiar Υ e que as quantidades de ruína podem tomar valores muito particulares (como se verá), a última probabilidade pode ser calculada com base nas expressões já vistas, dada pelo seguinte corolário [Drekic e Mera (2011)]:

Corolário para calcular $\boldsymbol{g}_{n-1,i-p_{k+n-1}}^{(k)}$

- se $\min\{c, \Upsilon + c_1\} = c$, então

$$\begin{cases} \boldsymbol{g}_{n-1,i-c}^{(k)} & \text{se } i = c, c+1, ..., \Upsilon + c_1 - 1 \\ \boldsymbol{g}_{n-1,i-c}^{(k)} + \sum\limits_{j=c_1}^{min\{c_2, i-\Upsilon\}} d_j \boldsymbol{g}_{n-1,i-j}^{(k)} & \text{se } i = \Upsilon + c_1, ..., min\{m_\alpha, \Upsilon + c\} - 1 \\ \sum\limits_{j=c_1}^{c_2} d_j \boldsymbol{g}_{n-1,i-j}^{(k)} & \text{se } i = \Upsilon + c, \Upsilon + c + 1, ..., m_\alpha - 1 \end{cases}$$

- se $\min\{c, \Upsilon + c_1\} = \Upsilon + c_1$, então

$$\begin{cases} \sum_{j=c_1}^{min\{c_2, i-\Upsilon\}} d_j \boldsymbol{g}_{n-1,i-j}^{(k)} & \text{se } i = \Upsilon + c_1, \Upsilon + c_1 + 1, ..., c-1 \\ \boldsymbol{g}_{n-1,i-c}^{(k)} + \sum_{j=c_1}^{min\{c_2, i-\Upsilon\}} d_j \boldsymbol{g}_{n-1,i-j}^{(k)} & \text{se } i = c, c+1, ..., min\{m_\alpha, \Upsilon + c\} - 1 \\ \sum_{j=c1}^{c2} d_j \boldsymbol{g}_{n-1,i-c}^{(k)} & \text{se } i = \Upsilon + c, \Upsilon + c + 1, ..., m_\alpha - 1 \end{cases}$$

Lembrando que $W_1 = k, \quad k \in \{1, 2, ..., n_r\}$, a Lei total das Probabilidades leva a que se conclua que

- $\phi_{n,j}(u) = \sum_{k=1}^{n_r} r_k \ \phi_{n-k,j}^{(k)}(u)$

- $\psi_{n,i,j}(u) = \sum_{k=1}^{n_r} r_k \ \psi_{n-k,i,j}^{(k)}(u)$
  $n = n_r + 1, n_r + 2, ...$

No entanto, pode acontecer que $T = n$ para $n = 1, 2, ..., n_r$, caso que não se deve esquecer. Há duas explicações possíveis para que aconteça:

1. a primeira indemnização ocorre em n^- para um limite de superávit de $u + cz_n + m$ (onde $\sum_{i=z_n}^{n-1} X_i = m$);

2. a primeira indemnização ocorre num instante $k^- \quad k \in \{1, 2, ..., n-1\}$ e a ruína ocorre $n - k$ unidades de tempo depois.

Combinando a informação anterior com a que resulta para os casos em que $n = n_r + 1, n_r + 2, ...$ $(n \in \mathbb{Z}^+)$, definem-se agora as expressões gerais para as quantidades de ruína que se pretendem calcular (tal como descrito no início da explicação do modelo):

- $\phi_{n,j}(u) = \sum_{k=1}^{min\{n-1, n_r\}} r_k \ \phi_{n-k,j}^{(k)}(u) + r_n \sum_{m=(n-z_n)c_1}^{(n-z_n)c_2} d_{m,n-z_n} \alpha_{u+cz_n+m+j}$;

- $\psi_{n,i,j}(u) = \sum_{k=1}^{min\{n-1, n_r\}} r_k \ \psi_{n-k,i,j}^{(k)}(u) + r_n \ d_{i-u-cz_n,n-z_n} \ \alpha_{i+j}$.

Notando que $\Lambda_{u+cn+m_\alpha-i} = 0$ se $m_\alpha = \infty$,

- a função de massa bivariada $\omega_{n,i}(u)$ é escrita como

$$\sum_{k=1}^{min\{n-1, n_r\}} r_k \sum_{j=1}^{m_\alpha - i} \psi_{n-k,i,j}^{(k)}(u) + r_n d_{i-u-cz_n,n-z_n} \left(\sum_{j=1}^{m_\alpha - i} \alpha_{i+j} \right)$$

$$= \Lambda_i \left(\sum_{k=1}^{min\{n-1, n_r\}} r_k (\boldsymbol{g}_{n-k-1,i-p_{n-1}}^{(k)} \boldsymbol{s}) + r_n d_{i-u-cz_n,n-z_n} \right).$$

Concluindo a explicação, é de referir que, para que seja possível analisar todo o processo, e uma vez que as expressões imediatamente acima resultam da soma das já obtidas para todos os valores de k unidades de tempo a considerar (e que dependerão obviamente de cada situação que se pretenda analisar), será necessário somar todos os valores que se pretendam estudar para o *défice na ruína* e *superávit imediatamente anterior à ruína*, obtendo assim a função de distribuição trivariada $\Psi_{n,x,y}(u)$, definida como

$$P(T \leq n, U_{T^-} \leq x, \mid U_T \mid \leq y \mid U_0 = u) = \sum_{l=1}^{n} \sum_{i=min\{c,Z+c_1\}}^{x} \sum_{j=1}^{y} \psi_{l,i,j}(u).$$

Este modelo foi inicialmente pensado como principal foco de trabalho, tendo sido iniciada a sua implementação. No entanto, devido às dificuldades enfrentadas nos processos de modelação e previsão de velocidade máxima de ventos, optou-se pela sua definição teórica, que, apesar de não ser original, salienta a capacidade de generalização associada à modelação de um processo tão complexo quanto o de montantes geridos por uma Seguradora, que se constitui como um dos processos de gestão mais lucrativos que se conhece.

Ilustração: Apresentam-se de seguida alguns resultados obtidos até ao momento pela implementação parcial deste modelo em R (com base em exemplos definidos em Drekic e Mera (2011)). Para o caso apresentado, assumem-se os seguintes **parâmetros**:

- f.d.p. geométrica truncada $a_j = \begin{cases} \left(\frac{2}{9}\right)\left(\frac{9}{11}\right)^{j-1} & \text{se } j = 1, 2, ..., n_a - 1 \\ \left(\frac{9}{11}\right)^{n_a - 1} & \text{se } j = n_a \end{cases}$ para $n_a = 10$
- f.d.p. pareto $\alpha_j = \left(1 + \frac{j-1}{30}\right)^{-4} - \left(1 + \frac{j}{30}\right)^{-4}$ para $j \in \mathbb{Z}^+$ e $m_a = 100$
- $d_i = \binom{5}{i}\left(\frac{2}{5}\right)^i\left(\frac{3}{5}\right)^{5-i}$, para $i = 0, ..., c$
- $c = 5$, $u = 2$, $\Upsilon = 50$

Apresenta-se de seguida uma parte do *output* da matriz P, definida por blocos:

```
name='scenario3'
Z=50
na=10
ma=1000
c=5
P=BIGP(na,ma,c,Z,name)
```

```
> P[1:10,1:10]
      [,1]          [,2]          [,3]          [,4]          [,5]          [,6]          [,7]          [,8]          [,9]          [,10]
 [1,] Numeric,100 Numeric,100 Numeric,100 Numeric,100 Numeric,100 ?           ?           ?           ?           ?
 [2,] Numeric,100 Numeric,100 Numeric,100 Numeric,100 Numeric,100 Numeric,100 ?           ?           ?           ?
 [3,] Numeric,100 Numeric,100 Numeric,100 Numeric,100 Numeric,100 Numeric,100 Numeric,100 ?           ?           ?
 [4,] Numeric,100 Numeric,100 Numeric,100 Numeric,100 Numeric,100 Numeric,100 Numeric,100 Numeric,100 ?           ?
 [5,] Numeric,100 Numeric,100 Numeric,100 Numeric,100 Numeric,100 Numeric,100 Numeric,100 Numeric,100 Numeric,100 ?
 [6,] Numeric,100 Numeric,100 Numeric,100 Numeric,100 Numeric,100 Numeric,100 Numeric,100 Numeric,100 Numeric,100 Numeric,100
 [7,] Numeric,100 Numeric,100 Numeric,100 Numeric,100 Numeric,100 Numeric,100 Numeric,100 Numeric,100 Numeric,100 Numeric,100
 [8,] Numeric,100 Numeric,100 Numeric,100 Numeric,100 Numeric,100 Numeric,100 Numeric,100 Numeric,100 Numeric,100 Numeric,100
 [9,] Numeric,100 Numeric,100 Numeric,100 Numeric,100 Numeric,100 Numeric,100 Numeric,100 Numeric,100 Numeric,100 Numeric,100
[10,] Numeric,100 Numeric,100 Numeric,100 Numeric,100 Numeric,100 Numeric,100 Numeric,100 Numeric,100 Numeric,100 Numeric,100
```

```
> P[1,1]
[[1]]
            [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10]
 [1,] 0.01206543    0    0    0    0    0    0    0    0     0
 [2,] 0.01206543    0    0    0    0    0    0    0    0     0
 [3,] 0.01206543    0    0    0    0    0    0    0    0     0
 [4,] 0.01206543    0    0    0    0    0    0    0    0     0
 [5,] 0.01206543    0    0    0    0    0    0    0    0     0
 [6,] 0.01206543    0    0    0    0    0    0    0    0     0
 [7,] 0.01206543    0    0    0    0    0    0    0    0     0
 [8,] 0.01206543    0    0    0    0    0    0    0    0     0
 [9,] 0.01206543    0    0    0    0    0    0    0    0     0
[10,] 0.06635989    0    0    0    0    0    0    0    0     0
```

```
> P[1,10]
[[1]]
10 x 10 sparse Matrix of class "dgCMatrix"

 [1,] 0 . . . . . . . . .
 [2,] . 0 . . . . . . . .
 [3,] . . 0 . . . . . . .
 [4,] . . . 0 . . . . . .
 [5,] . . . . 0 . . . . .
 [6,] . . . . . 0 . . . .
 [7,] . . . . . . 0 . . .
 [8,] . . . . . . . 0 . .
 [9,] . . . . . . . . 0 .
[10,] . . . . . . . . . 0
```

```
# 1. COMPUTE NON RUIN STATES -> NON RUIN STATES
C=P[(Z+1):(3*Z+1),(Z+1):(3*Z+1)]

# 2. COMPUTE NON RUIN STATES -> RUIN STATES
D=P[(Z+1):(3*Z+1),1:Z]
dimD1=dim(D)[1]
dimD2=dim(D)[2]
nb2=floor(dimD2/2)
j=1
while(j<=nb2){
  aux=D[,j]
  D[,j]=D[,dimD2-j+1]
  D[,dimD2-j+1]=aux
  j=j+1
}
```

Figura 5.2: P definida por Blocos e Extração de C e D - Exemplo Implementado.

Capítulo 6

Trabalho Futuro

Janssen e Manca (1997), que são os principais impulsionadores da aplicação de modelos semi-markovianos às mais variadas áreas, e que detêm diversas publicações sobre o tema, escrevem que o "Sob o ponto de vista computacional" (...) "é claro que este modelo não pode ser utilizado como um modelo de simulação sem ser na presença de uma boa máquina"[1]. Efetivamente, estes são modelos conceptualmente e computacionalmente "pesados", respetivamente devido à quantidade de informação de que dispõem e devido ao facto de terem, por norma, processos definidos por recorrência (como se pode ver na equação de evolução do processo semi-markoviano de primeira ordem). Os modelos semi-markovianos são uma generalização dos modelos de markov, motivo pelo qual dão uma maior liberdade na abordagem em qualquer tipo de problema que comporte modelação de processos que dependam do tempo, filas de espera, etc.

Como trabalho futuro seria interessante executar o mesmo tipo de simulação feita na modelação tendo em conta a recorrência temporal no processo de segunda ordem e avaliando as suas diferenças comportamentais em comparação com as restantes simulações. Seria também possível a consideração de variáveis explicativas no processo de modelação das velocidades de vento (por exemplo, as que se declararam significativas nos modelos de regressão), usando-as na estimação de P e F.

Relativamente à modelação da função densidade de probabilidade, e se se dispusesse de uma amostra temporal mais alargada, poder-se-ia recorrer a outros métodos de estimação da mesma, paramétricos ou não paramétricos, ou até uma conjugação dos dois (ver, por exemplo, McNeil (1999)). Do ponto de vista da previsão, uma análise mais aprofundada dos modelos ARIMA poderia sugerir melhores resultados a partir de um estudo temporal diferente dos mesmos dados (semanal, quinzenal, trimestral, mensal, avaliação por estações do ano, etc.). Note-se ainda que, para o tipo de problema considerado, dois anos de dados podem não ser suficientemente informativos sobre o processo, compromentendo a eficácia da previsão. Para além do trabalho efetuado, poder-se-iam considerar modelos não lineares para a previsão.

[1] tradução livre da autora.

Capítulo 7

Conclusões

Como é possível fazer uma previsão consistente quando se trata uma variável instável como o vento? A compreensão total deste tipo de processos ainda está longe de ter sido conseguida. A sua instabilidade e constante alteração de comportamento fazem com que sejam dos processos mais difíceis de prever e modelar, o que obriga à constante atualização dos métodos usados nestas tarefas.

Dada a dificuldade de conseguir bons resultados na previsão de velocidades máximas (horárias e diárias) de vento, esta tese acabou por se tornar num mecanismo de descoberta e exploração de métodos alternativos aos tradicionais. Uma boa parte do trabalho foi dedicada à modelação das funções densidade de probabilidade da variável em estudo, bem como à previsão por diversos métodos. Com efeito, os ensaios e tentativas de encontrar soluções satisfatórias foram-se sucedendo, ora encorajadores ora de alguma desilusão. Na sequência desse esforço, acabou por se optar pelo estudo dos valores diários na previsão, simplificando o problema. Concluiu-se que a abordagem linear ao problema pode não fornecer uma boa metodologia de tratamento do mesmo.

Quanto à modelação, os resultados foram um pouco mais satisfatórios, na medida em que foi explorado um processo de modelação com muito potencial, e que pode ser aplicado a muitas áreas para além da Climatologia (existindo diversas aplicações do mesmo em Seguros de Saúde, Seguros de Invalidez, Fundos de Pensões, Área da Banca, etc.). Em relação ao âmbito tratado nesta tese, mostrou-se que se trata de um modelo que apresenta melhores resultados quando comparado com o tradicional.

Bibliografia

Aigner T, Gjengedal T (2011). "Modelling wind power production based on numerical prediction models and wind speed measurements." *17th Power Systems Computation Conference, Stockholm*. (pág. 17)

Alfa AS, Drekic S (2007). "Algorithmic analysis of the Sparre Andersen model in discrete time." *Astin Bulletin*, **37**(02), 293–317. (pág. 63)

Asaduzzaman M, MahbubLatif A (2013). "A Markov Renewal Model for Predicting Tropical Cyclones in Bangladesh." (pág. 19)

Barbu V, Boussemart M, Limnios N (2004). "Discrete-time semi-Markov model for reliability and survival analysis." *Communications in Statistics-Theory and Methods*, **33**(11), 2833–2868. (pág. 19)

Barbu VS, Limnios N (2009). *Semi-Markov chains and hidden semi-Markov models toward applications: their use in reliability and DNA analysis*, volume 191. Springer Science & Business Media. (pág. 25)

Belsley DA, Kuh E, Welsch RE (1980). "Regression diagnostics: Identifying influential data and sources of collinearity." *Wiley Series in Probability and Mathematical Statistics, New York: Wiley, 1980*, **1**. (pág. 42)

Bessa RJ, Miranda V, Botterud A, Wang J, Constantinescu M (2012). "Time adaptive conditional kernel density estimation for wind power forecasting." *Sustainable Energy, IEEE Transactions on*, **3**(4), 660–669. (pág. 44)

Box GEP, Jenkins GM (1970). *Time Series Analysis: Forecasting and Control*. San Francisco: Holden-Day. (pág. 28)

Breiman L, Friedman J, Stone CJ, Olshen RA (1984). *Classification and regression trees*. CRC press. (pág. 34)

Brezinski C, Zaglia MR (2005). *Méthodes numériques directes de l'algèbre matricielle: cours et exercices corrigés, niveau L3*. Ellipses. (pág. 42)

Burges CJ (1998). "A tutorial on support vector machines for pattern recognition." *Data mining and knowledge discovery*, **2**(2), 121–167. (pág. 34)

Cabral MS, Gonçalves MH (2011). *Análise de dados longitudinais*. Sociedade Portuguesa de Estatística. (pág. 42)

Chi Yn, Liu Yh, Wang Ws, Chen Mz, Dai Hz (2007). "Study on Impact of Wind Power Integration on Power System [J]." *Power System Technology*, **3**, 77–8. (pág. 17)

D'Amico G, Petroni F, Prattico F (2012). "Reliability measures of second-order semi-Markov chain applied to wind energy production." *Journal of Renewable Energy*, **2013**. (pág. 49)

Davidson R, MacKinnon JG (1993). *Estimation and inference in econometrics*. Oxford University Press. (pág. 30)

Davidson R, MacKinnon JG (2004). *Econometric theory and methods*, volume 5. Oxford University Press New York. (pág. 30)

Dickey DA, Fuller WA (1979). "Distribution of the estimators for autoregressive time series with a unit root." *Journal of the American statistical association*, **74**(366a), 427–431. (pág. 30)

Drekic S, Mera AM (2011). "Ruin analysis of a threshold strategy in a discrete-time Sparre Andersen model." *Methodology and Computing in Applied Probability*, **13**(4), 723–747. (pág. 63), (pág. 69), (pág. 71)

Durbin J, Watson GS (1950). "Testing for serial correlation in least squares regression. I." *Biometrika*, **37**(3-4), 409–428. (pág. 42)

D'Amico G, Petroni F, Prattico F (2013). "First and second order semi-Markov chains for wind speed modeling." *Physica A: Statistical Mechanics and its Applications*, **392**(5), 1194–1201. (pág. 26), (pág. 51)

Elsner JB, Tsonis AA (2013). *Singular spectrum analysis: a new tool in time series analysis*. Springer Science & Business Media. (pág. 54)

Faraway JJ (2005). *Extending the linear model with R: generalized linear, mixed effects and nonparametric regression models*. CRC press. (pág. 33)

Faraway JJ (2009). *Linear models with R*. CRC Press. (pág. 40)

Golyandina N, Korobeynikov A (2014). "Basic singular spectrum analysis and forecasting with R." *Computational Statistics & Data Analysis*, **71**, 934–954. (pág. 54)

Golyandina N, Nekrutkin V, Zhigljavsky AA (2001). *Analysis of time series structure: SSA and related techniques*. CRC press. (pág. 54), (pág. 56)

Gonçalves E, Lopes Mendes N (2008). *Séries Temporais: Modelações Lineares e não lineares*. Sociedade Portuguesa de Estatística. (pág. 27), (pág. 29)

Green KC, Armstrong JS, Soon W (2009). "Validity of climate change forecasting for public policy decision making." *International Journal of Forecasting*, **25**(4), 826–832. (pág. 17)

Gybenko G (1989). "Approximation by superposition of sigmoidal functions." *Mathematics of Control, Signals and Systems*, **2**(4), 303–314. (pág. 33)

Haigh T (2008). "Cleve Moler: Mathematical software pioneer and creator of Matlab." *IEEE Annals of the History of Computing*, **1**(1), 87–91. (pág. 87)

Hand DJ, Mannila H, Smyth P (2001). *Principles of data mining*. MIT press. (pág. 31)

Hyndman RJ, Khandakar Y (2007). "Automatic time series for forecasting: the forecast package for R." *Relatório técnico*, Monash University, Department of Econometrics and Business Statistics. (pág. 52)

Iosifescu M, Limnios N, Oprisan G (2013). *Introduction to stochastic models*. John Wiley & Sons. (pág. 21)

Janssen J, Manca R (1997). "A realistic non-homogeneous stochastic pension fund model on scenario basis." *Scandinavian Actuarial Journal*, **1997**(2), 113–137. (pág. 73)

Janssen J, Manca R (2002). "General actuarial models in a semi-Markov environment." *Proceedings of ICA Cancun*, **2002**. (pág. 24)

Janssen J, Manca R (2006). *Applied semi-Markov processes*. Springer Science & Business Media. (pág. 21), (pág. 24)

Kwiatkowski D, Phillips PC, Schmidt P, Shin Y (1992). "Testing the null hypothesis of stationarity against the alternative of a unit root: How sure are we that economic time series have a unit root?" *Journal of econometrics*, **54**(1), 159–178. (pág. 30)

LeCun Y, Bottou L, Orr GB, Müller KR (1998). "Efficient BackProp." pp. 9–50. *Neural Networks: Tricks of the Trade, Springer*. (pág. 31)

Ljung GM, Box GE (1978). "On a measure of lack of fit in time series models." *Biometrika*, **65**(2), 297–303. (pág. 53)

Matloff N (2011). *The art of R programming: A tour of statistical software design*. No Starch Press. (pág. 87)

McNeil AJ (1999). "Extreme value theory for risk managers." *Departement Mathematik ETH Zentrum*. (pág. 73)

Moler C, *et al.* (1980). "Matlab user's guide." *Alberquerque, USA*. (pág. 87)

Nadaraya EA (1964). "On estimating regression." *Theory of Probability & Its Applications*, **9**(1), 141–142. (pág. 34)

Pinheiro C, Bates D (2000). *Statistics and Computing: Mixed-Effects Models in S and S-PLUS*. Springer, New York. (pág. 39)

Pyke R (1961). "Markov renewal processes: definitions and preliminary properties." *The Annals of Mathematical Statistics*, **32**, 1231–1242. (pág. 22)

R Core Team and others (2012). "R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing." (pág. 87)

Ross SM (2014). *Introduction to probability models*. Academic Press. (pág. 19)

Sansom J, Thomson P, *et al.* (2001). "Fitting hidden semi-Markov models to breakpoint rainfall data." *Journal of Applied Probability*, **38**, 142–157. (pág. 19)

Suykens JA, Vandewalle JP, de Moor BL (2012). *Artificial neural networks for modelling and control of non-linear systems*. Springer Science & Business Media. (pág. 33)

Torfs P, Brauer C (2014). "A (very) short introduction to R." *Hydrology and Quantitative Water Management Group Wageningen University*. (pág. 87)

Torgo L (2010). *Data mining with R: learning with case studies*. Chapman & Hall/CRC. (pág. 38), (pág. 87)

Vapnik VN (1995). *The nature of statistical learning theory*. Springer-Verlag New York, Inc. (pág. 34)

Watson GS (1964). "Smooth regression analysis." *Sankhyā: The Indian Journal of Statistics, Series A*, **26**, 359–372. (pág. 34)

Witten IH, Frank E (2005). *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann. (pág. 35)

Wu X, Li S (2008). *On a discrete-time Sparre Andersen model with phase-type claims*. Centre for Actuarial Studies: the research paper series, No. 169, University of Melbourne. (pág. 66)

Anexos

Anexo A

Demonstração. (referente à pagina 21). Seja $P_n(t) = P(N(t) = n)$. Considere-se

$$P_0(t+h) = P(N(t+h) = 0) = P(N(t) = 0, N(t+h) - N(t) = 0)$$

$$= P(N(t) = 0)P(N(t+h) - N(t) = 0) = P_0(t)(1 - \lambda h + o(h)).$$

Logo, obtém-se

$$P_0^{'}(t) = \lim_{h\to 0} \frac{P_0(t+h) - P_0(t)}{h} = \lim_{h\to 0} \Big(-\lambda P_0(t) + \frac{o(h)}{h}\Big) = -\lambda P_0(t).$$

A solução desta equação é $P_0(t) = Ce^{-\lambda t}$. Usando a condição inicial $P_0(0) = 1$, obtém-se $C = 1$ e conclui-se que $P_0(t) = e^{-\lambda t}$.
Para $n > 0$, obtém-se:

$$P_n(t+h) = P_n(t)P(N(t+h) - N(t) = 0) + P_{n-1}(t)P(N(t+h) - N(t) = 1) +$$

$$+ \sum_{k=2}^{n} P_{n-k}(t)P(N(t+h) - N(t) = k)$$

$$= P_n(t)(1 - \lambda h + o(h)) + P_{n-1}(t)(\lambda h + o(h)) + \sum_{k=2}^{n} P_{n-k}(t)o(h)$$

Logo

$$P_n^{'}(t) = \lim_{h\to 0} \frac{P_n(t+h) - P_n(t)}{h} = -\lambda P_n(t) + \lambda P_{n-1}(t) +$$

$$+ \lim_{h\to 0} \Big(P_n(t)\frac{o(h)}{h} + P_{n-1}(t)\frac{o(h)}{h} + \sum_{k=2}^{n} P_{n-k}(t)\frac{o(h)}{h}\Big) = -\lambda P_n(t) + \lambda P_{n-1}(t).$$

Reescrevendo a equação, tem-se

$$e^{\lambda t}(P_n^{'}(t) + \lambda P_n(t)) = \lambda e^{\lambda t} P_{n-1}(t) \quad \text{ou} \quad \frac{d}{dt}(e^{\lambda t} P_n(t)) = \lambda e^{\lambda t} P_{n-1}(t).$$

Assim,

$$\frac{d}{dt}(e^{\lambda t} P_n(t)) = \frac{\lambda^n t^{n-1}}{(n-1)!} \quad \text{ou} \quad e^{\lambda t} P_n(t) = \frac{(\lambda t)^n}{n!} + C.$$

Pela condição inicial $C = 0$, logo $P_n(t) = e^{-\lambda t}\frac{(\lambda t)^n}{n!}$. □

Informação de apoio à decisão

Aqui estão incluídas algumas análises gráficas ou numéricas de *outputs* das diferentes instruções usadas em R ao longo desta tese. Esta informação é referida de modo contextualizado, referente ao Capítulo de Aplicação:

- **CAPÍTULO 4 - APLICAÇÃO**:
  - Resultado da aplicação do método dos mínimos generalizados no caso horário da primeira etapa de análise da base de dados:

```
Generalized least squares fit by REML
Model: VMax ~ I(A) + I(M) + I(H) + I(E) + TMin + HMax + RMin + DV + PT
Data: BaseDados
AIC      BIC      logLik
92031.63 92132.81 -46002.81

Coefficients:
                 Value    Std.Error   t-value  p-value
(Intercept)   -193.63100  94.41517  -2.05085  0.0403
I(A)             0.10226   0.04686   2.18209  0.0291
I(M)             0.02959   0.01328   2.22820  0.0259
I(H)             0.00537   0.00356   1.51173  0.1306
I(E)Outono       0.49885   0.12423   4.01566  0.0001
I(E)Primavera   -0.79088   0.08837  -8.94990  0.0000
I(E)Verão       -0.30683   0.12943  -2.37060  0.0178
TMin            -0.18092   0.00716 -25.28111  0.0000
HMax            -0.03994   0.00152 -26.30061  0.0000
RMin             0.00220   0.00015  14.27522  0.0000
DV              -0.01349   0.00022 -61.81387  0.0000
PT               0.46016   0.02432  18.92248  0.0000
```

  - Resultado da aplicação da regressão linear no caso diário da primeira etapa de análise da base de dados:

```
Call:
lm(formula = VMax ~ I(E) + TMed +
HMed + DV + PT,data = d)

Residuals:
Min      1Q  Median      3Q     Max
-9.3278 -2.3320 -0.3839  2.0254 14.5806

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)               18.825446   0.679681  27.697  < 2e-16 ***
I(d$Estacao)Outono         1.269779   0.426506   2.977 0.003007 **
I(d$Estacao)Primavera      0.237417   0.460728   0.515 0.606496
I(d$Estacao)Verão          1.818327   0.612668   2.968 0.003098 **
d$Tmed                    -0.403030   0.034928 -11.539  < 2e-16 ***
d$HMed                    -0.031680   0.008215  -3.856 0.000125 ***
d$DV                      -0.014155   0.001664  -8.505  < 2e-16 ***
d$PT                       0.169784   0.071727   2.367 0.018192 *
```

```
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.488 on 720 degrees of freedom
Multiple R-squared:  0.4042,Adjusted R-squared:  0.3984
F-statistic: 69.77 on 7 and 720 DF,  p-value: < 2.2e-16
```

– Resultado da aplicação das funções de modelação de variância - melhor modelo obtido:

```
Generalized least squares fit by REML
Model: VMax ~ I(A) + I(M) + I(Estacao) + TMin + HMax + RMin + DV + PT
Data: BaseDados
AIC      BIC   logLik
88984.99 89109.52 -44476.5

Variance function:
Structure: Exponential of variance covariate, different strata
Formula: ~TMin | E
Parameter estimates:
Verão      Outono     Inverno   Primavera
-0.04943755 -0.03987934 -0.05740226 -0.05254327

Coefficients:
Value Std.Error   t-value p-value
(Intercept)         -620.2921  77.89434  -7.96325  0.0000
I(A)                   0.3139   0.03866   8.11979  0.0000
I(M)                   0.0840   0.01374   6.11193  0.0000
I(Estacao)Outono       0.5408   0.12735   4.24620  0.0000
I(Estacao)Primavera   -0.7176   0.08818  -8.13741  0.0000
I(Estacao)Verão       -0.3545   0.11579  -3.06165  0.0022
TMin                  -0.1638   0.00631 -25.94540  0.0000
HMax                  -0.0491   0.00139 -35.30828  0.0000
RMin                   0.0018   0.00010  17.51989  0.0000
DV                    -0.0106   0.00019 -54.70199  0.0000
PT                     0.4684   0.02762  16.95941  0.0000

Correlation:
             (Intr)  I(A)   I(M)  I(Es)O  I(Es)P I(Es)V  TMin  HMax   RMin   DV
I(A)         -1.000
I(M)         -0.251  0.250
I(Estacao)O  -0.057  0.057 -0.678
I(Estacao)P   0.150 -0.150 -0.236  0.608
I(Estacao)V  -0.132  0.133 -0.395  0.763  0.779
TMin         -0.036  0.035 -0.091 -0.245 -0.536 -0.631
HMax          0.049 -0.050  0.025 -0.205 -0.221 -0.228  0.215
RMin         -0.037  0.037  0.079  0.058  0.098  0.164 -0.437  0.258
DV           -0.027  0.028 -0.017  0.062  0.029  0.015 -0.042 -0.280  0.015
PT           -0.008  0.008 -0.046  0.033  0.032  0.027  0.054 -0.120 -0.040  0.037

Standardized residuals:
Min         Q1        Med         Q3        Max
-3.9028458 -0.6491694 -0.1364726  0.5777264  6.6580570

Residual standard error: 6.743717
```

Seguidamente são apresentados os resultados da previsão temporal (horária) por uso do modelo linear ARMA (com base na transformação de dados e determinação da componente sazonal do modelo) e por uso de redes neuronais artificiais, com 1 camada escondida (e divisão de dados em 70% para teste, 15% para treino e 15% para validação). Devido à alta correlação residual, optou-se pela colocação destes resultados apenas para visualização. Note-se que a componente sazonal parece ter um bom ajuste aos dados, mas foi estimada com base numa série harmónica (combinação linear de senos e cossenos) por regressão linear. Ora, já foi visto que a regressão linear não deve ser considerada nestes dados, uma vez que não apresentam independência.

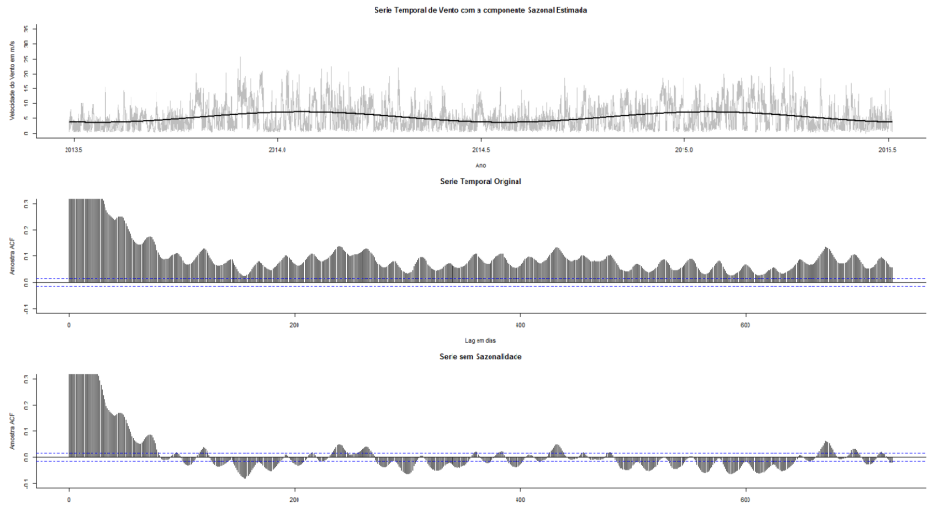


Figura 1: Modelo Arima em Dados Horários.

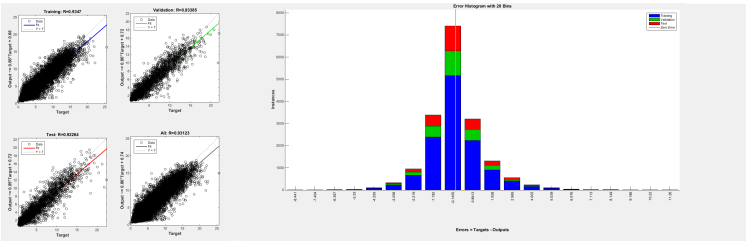


Figura 2: Redes Neuronais Artificiais em Dados Horários.

Anexo B

Ferramentas de trabalho

Para a exploração da base de dados e implementação dos modelos já referidos, foram utilizadas duas ferramentas de programação: R e MATLAB.

O R [R Core Team and others (2012)] é um ambiente de desenvolvimento integrado orientado a objetos, direcionado sobretudo para análise e manipulação de dados. Entre as suas principais vantagens, destacam-se o facto de ser uma aplicação de distribuição gratuita e de código público, existindo versões já compiladas para os principais sistemas operativos, que, na sua maioria, são fornecidas pela comunidade de utilizadores. É particularmente útil para lidar com grandes conjuntos de dados [Torgo (2010)] e apresenta bons tempos de execução nessas circunstâncias [Matloff (2011)]. No entanto, existe uma desvantagem de relevo: a existência de *packages* pré-implementados ou definidos pelos utilizadores - e disponíveis para os restantes - obriga ao conhecimento da teoria que está por trás das implementações, para que se possa fazer pleno uso desses *packages*. O que se pretende dizer com isto é que o uso de pré-implementações não deve ser feito de forma despreocupada e sem tentar perceber qual o raciocínio que lhes está associado, quando este não é óbvio. O código fonte para o ambiente de software é escrito principalmente em C, FORTRAN e R.

Para uma programação mais intuitiva neste ambiente, optou-se pela utilização do RStudio [Torfs e Brauer (2014)], que é um ambiente de desenvolvimento integrado (IDE) para o R e que, por esse motivo, apresenta características e ferramentas de apoio ao desenvolvimento de software com o objetivo de agilizar este processo (como *autocomplete* de comandos, indicações gráficas de fecho e abertura de parêntesis, indicações de nomes internos de funções, disponibilização fácil do código das intruções, etc.).

O MATLAB (diminutivo de MATrix LABoratory), é um software interativo particularmente indicado para o cálculo numérico e usado desde a década de 70 [Moler *et al.* (1980), Haigh (2008)], que integra análise numérica, cálculo matricial, métodos de processamento de sinais e construção gráfica. Esta ferramenta foi utilizada como validação de alguns resultados obtidos, tendo sido usadas as *toolbox* de Redes Neuronais e a aplicação Semi-Markov. Esta última foi utilizada para poder comparar os resultados provenientes do algoritmo implementado pela autora desta tese e os resultados provenientes dessa *toolbox*, vistos no gráfico da esquerda da Figura 4.5.

Anexo C

Categorização da velocidade de vento

Existe mais do que uma escala de categorização de vento, entre as quais a **escala Beaufort**, que é um sistema que relaciona a velocidade do vento com as condições observadas no mar ou em terra, isto é, com os efeitos da mesma sobre estes dois meios físicos.

Os critérios estão descritos nas Tabelas 1 e 2 e foram baseados na informação que se pode encontrar no site da Organização Mundial de Meteorologia `https://www.wmo.int/pages/index_en.html` [1]. Note-se que na base de dados em estudo, na estação meteorológica de Settala, cuja localização geográfica se mostra em seguida, o valor máximo de velocidade de vento é de 25 m/s, não se justificando a consideração de intensidades de vento elevadas. No entanto, ponderou-se a discretização (separação de valores por estado) aliada a um critério climatológico, também descrito abaixo, na Tabela 3.

Intensidade	**Descrição**	**mph**	**m/s**
0	Calmo	< 1	< 0.3
1	Aragem	1 - 3	0.3 - 1.5
2	Brisa Leve	4 - 7	1.6 - 3.3
3	Brisa Fraca	8 - 12	3.4 - 5.4
4	Brisa Moderada	13 - 18	5.5 - 7.9
5	Brisa Forte	19 - 24	8 - 10.7
6	Vento Suave	25 - 31	10.8 - 13.8
7	Vento Forte	32 - 38	13.9 - 17.1
8	Ventania	39 - 46	17.2 - 20.7
9	Ventania Forte	47 - 54	20.8 - 24
10	Tempestade	55 - 63	24.5 - 28.4
11	Tempestade Violenta	64 - 74	28.5 - 32.6
12	Furacão	≥ 74	≥ 32.7

Tabela 1: Escala de Beaufort - Intensidades de vento.

[1] última consulta realizada em 30 de Agosto de 2015.

Intensidade	**Efeitos no Mar (M) e em Terra (T)**
0	efeito espelhado fumo ascendente
1	pequenas ondulações aparecimento de rastos de fumo
2	ondulação sem rebentação movimentação de moinhos
3	ondulação até 60 cm desfraldar de bandeiras
4	ondulação até 1 m movimentação de galhos das árvores
5	ondulação até 2.5 m movimentação de galhos e árvores pequenas
6	grandes ondas até 3.5 m movimentação dos ramos das árvores
7	mar revolto até 4.5 m movimentação de grandes árvores
8	mar revolto até 5 m com rebentação quebra de galhos de árvores
9	mar revolto até 7 m Danos em árvores e pequenas construções
10	mar revolto até 9 m árvores arrancadas e danos estruturais em construções
11	mar revolto até 11 m estragos generalizados em cosntruções
12	mar revolto até 14 m, sem visibilidade estragos graves e generalizados em construções

Tabela 2: Escala de Beaufort - Efeitos da velocidade do vento.

Estado	Velocidade de Vento (em m/s)
1	< 4.27
2	4.27 - 8.53
3	8.53 - 12.8
4	12.8 - 17.1
5	17.1 - 21.3
6	≥ 21.3

Tabela 3: Critério de Discretização da Variável Velocidade Máxima de Vento.

Contextualização geográfica do problema

Settala é uma comuna italiana da região da Lombardia, província de Milão, com cerca de 5790 habitantes. Estende-se por uma área de 17 km^2, tendo uma densidade populacional de 341 hab/km^2. Seguidamente é mostrado um mapa que fornece a localização geográfica da estação meteorológica onde foram realizadas as medições presentes na base de dados.

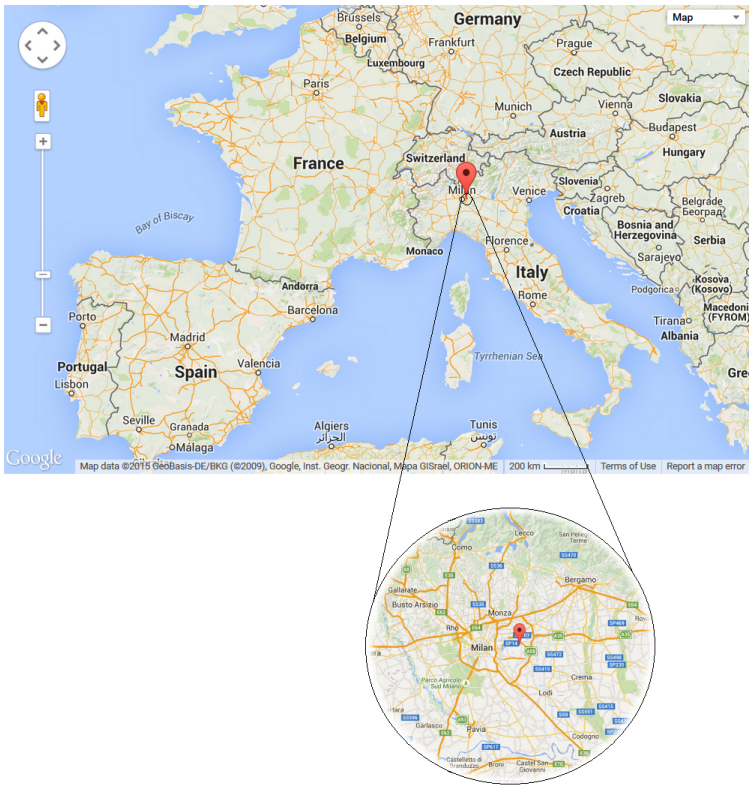


Figura 3: Mapa da Localização da Estação Meteorológica de Settala.

Anexo D

Exploração da Base de Dados

Este anexo é usado para avaliar, uma a uma, as 14 variáveis não temporais numéricas e a variável sazonal categórica presentes na versão tratada da base de dados, fazendo um sumário das características das mesmas. Como as unidades de medida das diferentes variáveis climáticas são diferentes, foi realizada uma normalização das mesmas, cujos diagramas de caixa e bigodes (vulgo *boxplot*) se mostram em seguida:

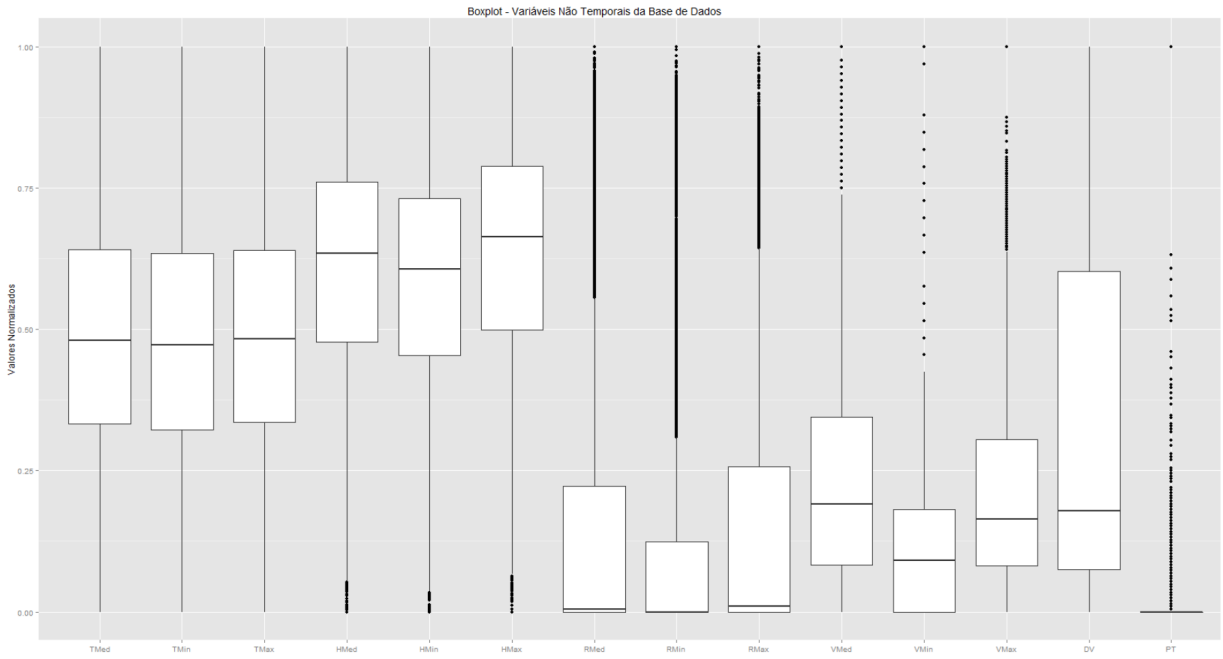
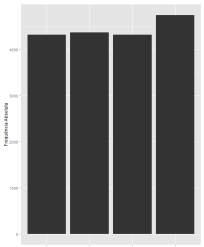


Figura 4: Boxplot dos Dados Normalizados.

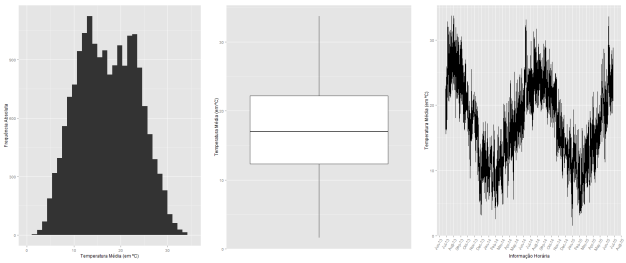
É possível observar que as variáveis associadas com a Humidade Relativa, Radiação Global, Velocidade de Vento e Precipitação Total apresentam *outliers*. As variáveis associadas à Radiação Global aparentam ser assimétricas negativas (apresentam enviesamento à esquerda), ao passo que as restantes referidas na frase anterior aparentam ser assimétricas positivas (enviesadas à direita).

Seguidamente é feita uma descrição gráfica e numérica das variáveis da base de dados.

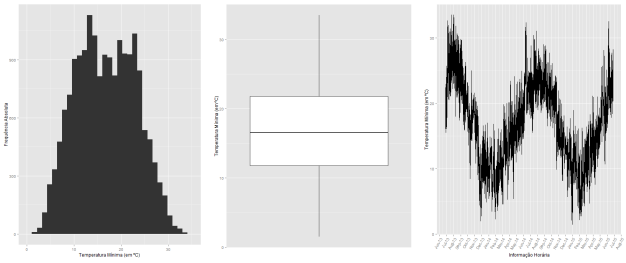
E: Estação do Ano

Descrição Numérica	**Descrição Gráfica**
• Inverno: 4319 observações • Outono: 4366 observações • Primavera: 4319 observações • Verão: 4739 observações	

TMed: Temperatura Média

Descrição Numérica	**Descrição Gráfica**
• Mínimo amostral: 1.6 • Média amostral: 17.18 • Máximo amostral: 33.7 • 1º Quartil: 12.3 • 2º Quartil: 17 • 3º Quartil: 22.2	

TMin: Temperatura Mínima

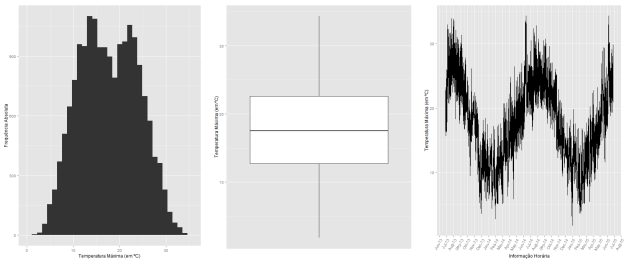
Descrição Numérica	**Descrição Gráfica**
• Mínimo amostral: 1.5 • Média amostral: 17.18 • Máximo amostral: 33.5 • 1º Quartil: 11.8 • 2º Quartil: 16.6 • 3º Quartil: 21.8	

TMax: Temperatura Máxima

Descrição Numérica

- Mínimo amostral: 1.8
- Média amostral: 17.67
- Máximo amostral: 34.3
- 1º Quartil: 12.7
- 2º Quartil: 17.5
- 3º Quartil: 22.6

Descrição Gráfica

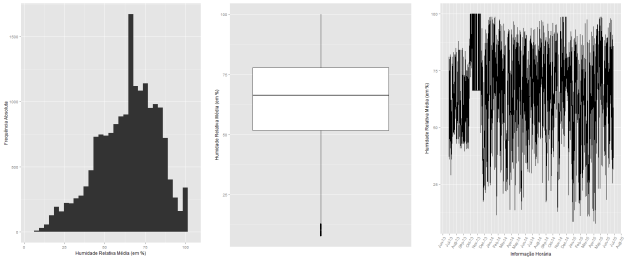


HMed: Humidade Relativa Média

Descrição Numérica

- Mínimo amostral: 7.8
- Média amostral: 64.16
- Máximo amostral: 100.0
- 1º Quartil: 51.8
- 2º Quartil: 66.3
- 3º Quartil: 77.9

Descrição Gráfica

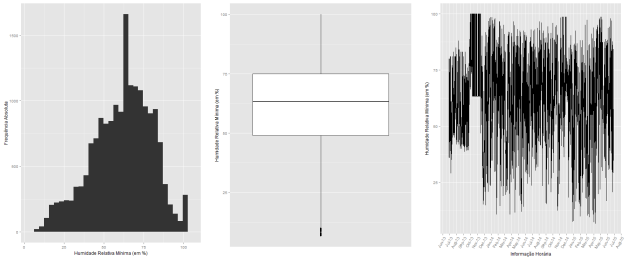


HMin: Humidade Relativa Mínima

Descrição Numérica

- Mínimo amostral: 6.7
- Média amostral: 61.33
- Máximo amostral: 100.0
- 1º Quartil: 49
- 2º Quartil: 63.6
- 3º Quartil: 75

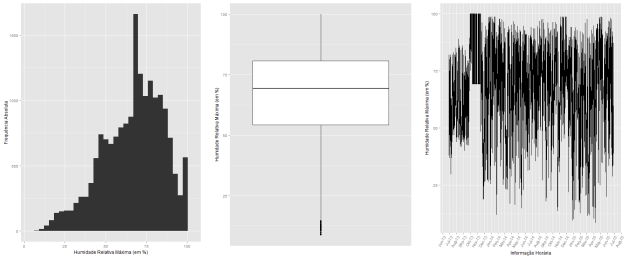
Descrição Gráfica



HMax: Humidade Relativa Máxima

Descrição Numérica **Descrição Gráfica**

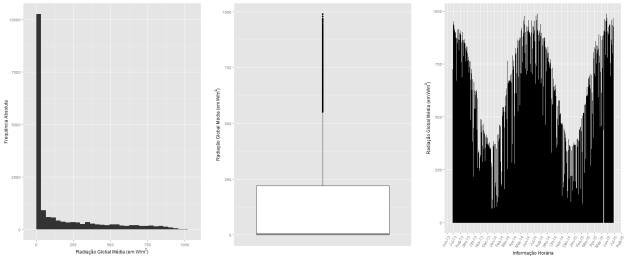
- Mínimo amostral: 8.7
- Média amostral: 66.93
- Máximo amostral: 100.0
- 1º Quartil: 54.3
- 2º Quartil: 69.3
- 3º Quartil: 80.7



RMed: Radiação Global Média

Descrição Numérica **Descrição Gráfica**

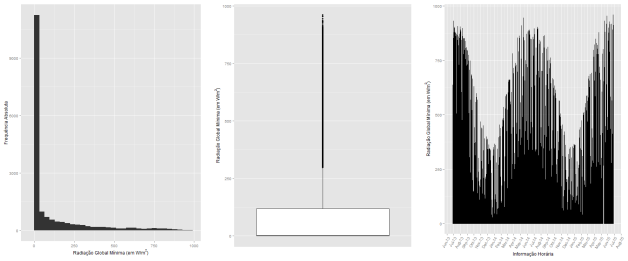
- Mínimo amostral: 0.0
- Média amostral: 149.1
- Máximo amostral: 988.0
- 1º Quartil: 0.0
- 2º Quartil: 4.4
- 3º Quartil: 220



RMin: Radiação Global Mínima

Descrição Numérica **Descrição Gráfica**

- Mínimo amostral: 0.0
- Média amostral: 103.5
- Máximo amostral: 958.9
- 1º Quartil: 0.0
- 2º Quartil: 0.0
- 3º Quartil: 118.6

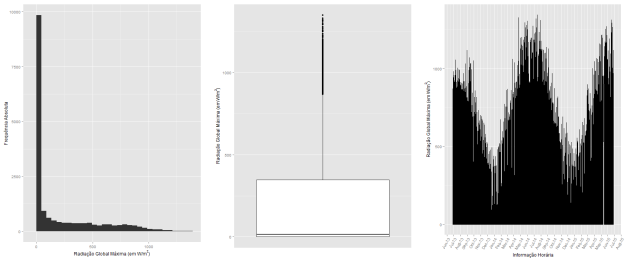


RMax: Radiação Global Máxima

Descrição Numérica

- Mínimo amostral: 0.0
- Média amostral: 203.31
- Máximo amostral: 988.0
- 1º Quartil: 0.1
- 2º Quartil: 14.1
- 3º Quartil: 347

Descrição Gráfica

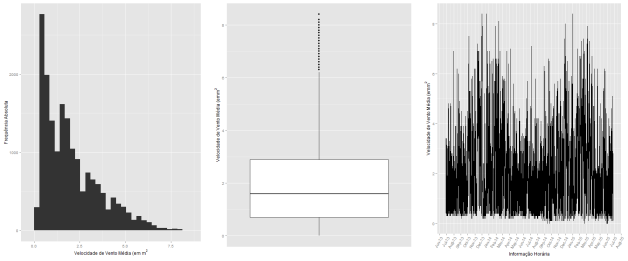


VMed: Velocidade de Vento Média

Descrição Numérica

- Mínimo amostral: 0.0
- Média amostral: 2.016
- Máximo amostral: 8.4
- 1º Quartil: 0.7
- 2º Quartil: 1.6
- 3º Quartil: 2.9

Descrição Gráfica



VMin: Velocidade de Vento Mínima

Descrição Numérica

- Mínimo amostral: 0.0
- Média amostral: 0.3976
- Máximo amostral: 3.3
- 1º Quartil: 0.0.0
- 2º Quartil: 0.3
- 3º Quartil: 0.6

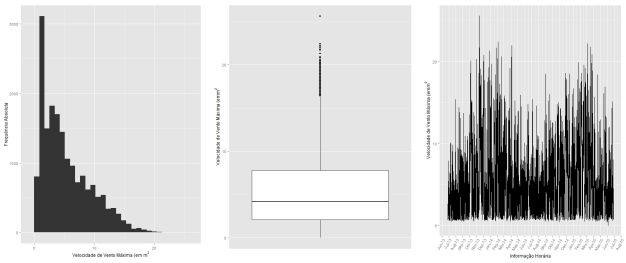
Descrição Gráfica

VMax: Velocidade de Vento Máxima

Descrição Numérica

- Mínimo amostral: 0.0
- Média amostral: 5.35
- Máximo amostral: 25.6
- 1º Quartil: 2.1
- 2º Quartil: 4.2
- 3º Quartil: 7.8

Descrição Gráfica



Seguidamente são consideradas duas medidas de inferência estatística úteis para fazer considerações teóricas sobre a variável que está a ser estudada (seja X essa variável, onde $X = x_1, x_2, ..., x_n$ é a amostra analisada):

- **Coeficiente de Curtose** dado por

$$\gamma_2 = \frac{\mu_4}{\mu_2^2} - 3 = \frac{E[(X - E[X])^4]}{(E[(X - E[X])^2])^2} - 3,$$

  onde μ_2 e μ_4 são, respetivamente, os momentos centrados de segunda e quarta ordens. Caso $\gamma_2 > 0$, a distribuição da variável é apelidada de leptocúrtica; caso $\gamma_2 < 0$, diz-se platicúrtica e apresenta uma distrubuição plana dos dados; quando $\gamma_2 = 0$, diz-se mesocúrtica e deve ter um comportamento semelhante a uma população distribuída segundo uma Normal.

- **Coeficiente de Assimetria** dado por

$$\gamma_1 = \frac{\mu_3}{\mu_2^{(\frac{3}{2})}} = \frac{E[(X - E[X])^3]}{(E[(X - E[X])^2])^{\frac{3}{2}}}.$$

  $\gamma_1 > 0$ significa que a população em análise apresenta uma média superior à mediana, denominando-se como assimétrica à direita. Caso contrário, é assimétrica à esquerda e apresenta maior mediana do que média. μ_3 é o momento centrado de terceira ordem.

Variável	TMed	TMin	TMax	HMed	HMin	HMax	RMed
γ_1	0.054	0.056	0.055	−0.449	−0.382	−0.499	1.636
γ_2	−0.821	−0.819	−0.812	−0.219	−0.240	−0.191	1.542

Variável	RMin	RMax	VMed	VMin	VMax
γ_1	2.224	1.423	1.086	1.684	0.987
γ_2	4.38	0.868	0.677	3.323	0.359